Test Mathematical Statistics (M5-201800139),
November 8, 2018, 8.45-11.45 h.
Instructors Johannes Schmidt-Hieber and Dick Meijer, module coordinator
Pranab Mandal

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator ("GR") is not. No tables of probability distributions are needed for the test.

Part A: Basic concepts

(a) [1 point] What is the difference between categorical and quantitative/numerical variables.

(b) [1 point] What is the difference between the coefficient of determination R^2 and the adjusted R^2.

(c) [1 point] Describe the difference between Fisher's exact test and the χ^2-test.

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

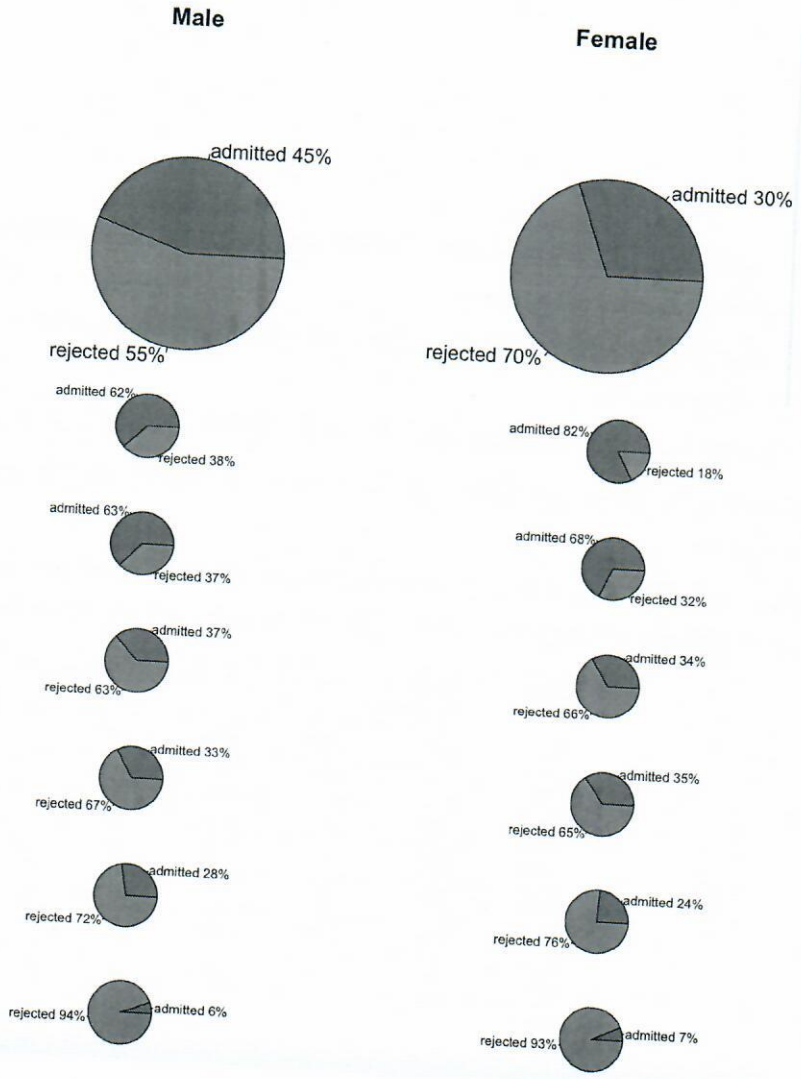
(e) [1 point] There is a debate whether some of Jane Austen's novels might have been written by a ghost writer. The table below gives some word frequencies.

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 in order to check whether a ghostwriter has been involved.

Part B: Visual interpretation of data

1. [2 points] Below are pie charts of the percentages of admission/rejection for male and female applicants at the University of Berkeley in 1973. The larger pie charts in the first row show the percentages for the whole university. The smaller pie charts show the percentages of admission/rejection for each of the six individual faculties. While for the data from the whole university, there seems to be a gender bias with female applicants being much more frequently rejected, on the level of individual faculties there is little difference between the genders. Give a statistical explanation why there is no contradiction.



2. [1 point] Sketch a scatterplot in which the correlation without an outlier is positive, but the correlation when the outlier is added is negative. Indicate in your plot which point is the outlier.

3. [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).

Part C: Theory

4. 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?

5. Suppose that in a random sample of 200 students who did not take the mathematical statistics course yet, 20% say that they like statistics. In a second random sample of 200 students who recently took the statistics course, 80% like statistics. Is this a significant increase?

   (a) [2 points] State the null and alternative hypothesis and say which statistical test should be applied here.

   (b) [3 points] Derive the test statistic for this testing problem. Can we reject the null hypothesis at level $\alpha = 0.025$? *Remark: It is fine to do the computations without continuity correction.*

   Hint: For the computations, you [redacted] should use approximation by a normal distribution. You can use that $\Phi^{-1}(0.025) = -\Phi^{-1}(0.975) \approx -2$.

6. 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 $Y - X\widehat{\beta}^{\text{MLE}}$.

   (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?

7. [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.

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: $\quad var(X+Y) = var(X) + var(Y), \quad var(X-Y) = var(X) + var(Y)$

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

Distribution	Probability/Density function	Range	$E(X)$	$var(X)$
Binomial (n,p)	$\binom{n}{x} p^x (1-p)^{n-x}$	$0, 1, 2, \ldots, n$	np	$np(1-p)$
Poisson (μ)	$e^{-\mu}\mu^x/x!$	$0, 1, 2, \ldots$	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^2_f)	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 $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x}$, $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \overline{y})$, $\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$, $se(\hat{\beta}_1) = \frac{S}{\sqrt{S_{xx}}}$ 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)
* $T = \dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$ and S^2
* $T = \dfrac{(\overline{X}-\overline{Y})-\Delta_0}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, with $S^2 = \dfrac{n_1-1}{n_1+n_2-2}S_X^2 + \dfrac{n_2-1}{n_1+n_2-2}S_Y^2$ or: $Z = \dfrac{(\overline{X}-\overline{Y})-\Delta_0}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}}$
* $F = \dfrac{S_X^2}{S_Y^2}$
* $Z = \dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, with $\hat{p} = \dfrac{X_1+X_2}{n_1+n_2}$
* (regression) $T = \hat{\beta}_i / se(\hat{\beta}_i)$ and $F = \dfrac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$

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

Analysis of categorical variables

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