THE GRADUATE INSTITUTE GENEVA

INSTITUT DE HAUTES
ÉTUDES INTERNATIONALES
ET DU DÉVELOPPEMENT

GRADUATE INSTITUTE
OF INTERNATIONAL AND
DEVELOPMENT STUDIES

Statistics for International Relations Research I

Waiver Test

Rémi Viné

September, 2022
The Graduate Institute — Geneva

Instructions

- You have **90 minutes to complete the test**. The total number of points assigned is 90.
- You can solve exercises in the order you prefer, but you should make it clear(by labelling exercises and questions on your booklet).
- Please write the results (and perform computations) with 4 decimal digits.
- Show all your work in the answer booklet, you will not receive full credit if you only write the answer.
- This is an individual exercise, do not interact with other students.
- You can use a calculator, except for the one on your smartphone.
- You are given a formula sheet.
- You are given tables for the standard normal distribution and the Student's t-distribution.

1 Exercise *(8 points)*

For each of the following random variables, determine (i) whether the variable is categorical or numerical, (ii) if the variable is numerical, indicate whether it is discrete or continuous, (iii) the level of measurement (nominal, ordinal, interval, ratio)

1. The exchange rate between the Ruble and the CHF (Swiss Franc)
2. The temperature of the day in Celsius degree
3. Ingredients in a pizza
4. Number of passengers in the tram line #15 per week

2 Exercise *(8 points)*

A Pew Research Center poll analyzed Indian public opinion on national conditions in 2018. It is stated that ”About two-thirds of Indians (66%) believe that today’s children will be better off than their parents. But that optimism is down 10 percentage points since 2017.”

1. Is this statement descriptive or inferential?
2. How to rephrase it so that it becomes descriptive (if you replied that it is inferential in 1)) or inferential (if you replied descriptive in 1))?
3. What was the share of optimists in 2017?
4. If, instead of being ”down by 10 percentage points since 2017”, optimism were down by 10%, what would have been the 2017 share of optimists?

3 Exercise *(5 points)*

1. Can you find the standard deviation and the inter-quartile range from the following R output? Give the value(s) of the statistics measuring spread that can be retrieved.

```
p = seq(0,1, length=1000)
beta_distr <- dbeta(p, 1, 8)
summary(beta_distr)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000488 0.062501 1.003005 1.067879 8.000000
```

2. In a distribution where the mean is larger than the median, what type of asymmetry of the distribution would you expect to observe? *You may draw a distribution to illustrate.*

4 Exercise *(10 points)*

	1	2	3	4
x_i	2	23	45	51
$P(X = x_i)$	0.3	0.1	0.2	0.4

Table 1: Probability distribution for a random variable X, the (fictitious) price to pay for a train ticket Geneva - Bern, in CHF

1. Find the expected value, $E(X)$. Interpret.
2. What is the median? Interpret and comment on the potential skewness of the distribution.
3. Assume the SBB (Swiss train company) decides to stop the super cheap tickets (CHF 2) and rather propose to those paying CHF 2 to now pay CHF 23. If there is no change in preferences (no substitution with other means of transportation so that all the CHF 2 buyers are shifted to the CHF 23 buyers), would the expected value increase or decrease? And what about the standard deviation? (You are not asked to compute the new statistics, you are simply asked to provide the intuition backing your responses)

5 Exercise *(11 points)*

The table, from the IDMC 2022 report, indicates the number of internally displaced young people by region. Probabilities are asked assuming one could select a person out of the universe of all internally displaced young people. You may use the given acronyms to facilitate the notation.

	0-4yo	5-14yo	15-24 yo
Sub-Saharan Africa (SSA)	4.4	7.3	5.5
Middle East and North Africa (MENA)	1.5	2.7	5
South Asia (SA)	0.9	1.8	1.5
Americas (A)	0.5	1	1.1
Europe and Central Asia (ECA)	0.2	0.5	0.4
East Asia and Pacific (EAP)	0.2	0.5	0.4

Table 2: Age groups and regions of origin of internally displaced young people (in million)

1. What is the probability for an internally displaced young person to come from the Americas?

2. What is the probability for an internally displaced young person to be aged 15-24 years old or come from "East Asia and Pacific"?

3. Given that the young internally displaced person is aged 0-4 years old, what is the probability that this person comes from Sub-Saharan Africa?

4. Are being 0-4 years old and coming from Sub-Saharan Africa independent events? Justify and interpret the answer to the question.

6 Exercise *(11 points)*

1. Figure 1 shows the standard normal distribution and the student distribution with 1 degree of freedom. Which one is the standard normal distribution? Justify.

2. What is the area under the curve on the right of the right vertical line (x = 1.96) for the standard normal distribution?

3. What is the area under the standard normal curve that is between the left and the right vertical lines ($x = -2.33$ and $x = 1.96$)?

4. In what context should the student distribution be used for hypothesis testing and confidence interval construction as compared to the normal distribution? What is the theorem, typical in statistics, that is underlying this?

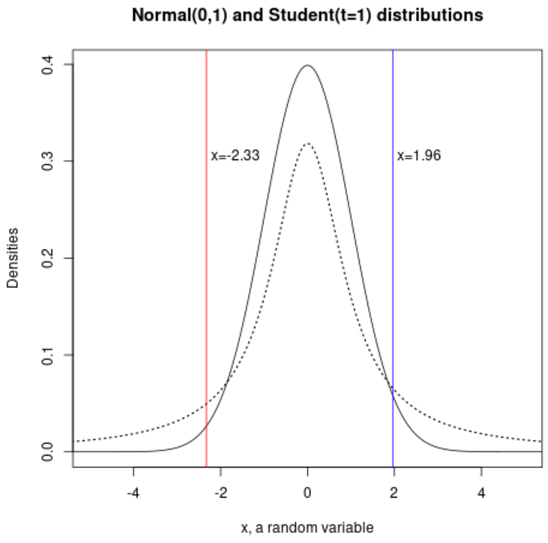


Figure 1: Normal and Student distributions

7 Exercise *(14 points)*

A BBC world service opinion poll surveyed 30352 people on climate change perceptions. It is found that 56% of the respondents want their governments to be more proactive in addressing climate change challenges. (To simplify, we assume that the sampling was random.)

1. Build a 99.2 % confidence interval for the proportion of respondents who want their governments to be more proactive toward climate change. Interpret it.

2. What is the point estimate, and what is the standard error of the confidence interval?

3. Assume that the BBC service has to reduce costs and therefore surveys "only" 5000 people. How would it change the 99.2 % confidence interval (and what is changing: point estimate, margin of error or t-statistic)? You may not need to compute the new confidence interval.

8 Exercise *(23 points)*

Note on the data used. A data frame with X observations on 6 variables, each of which is in percent, i.e., in [0,100]. The data characterizes some French-speaking provinces of Switzerland in 1888.

Fertility	I_g , 'common standardized fertility measure'
Agriculture	% of males involved in agriculture as occupation
Examination	% draftees receiving highest mark on army examination
Education	% education beyond primary school for draftees
Catholic	% 'catholic' (as opposed to 'protestant')
Infant.Mortality	live births who live less than 1 year

Table 3: Variables - presentation

1. Among the variables displayed in the summary table, what variable is the most skewed?

2. Unfortunately, the summary statistics table fails to display the number of observations, can you retrieve the number of observations using the regression output (explain how)?

3. Interpret the intercept. Does it bring meaningful information?

4. Interpret the slope coefficient of education. Would you interpret it as *education causes low/higher fertility*?

5. In a province with 25% males involved in agriculture, 30% draftees receiving highest mark on army examination, 50% of draftees beyond primary education, 80% of Catholics, and 15% of live births who live less than one year, what would be the expected fertility index?

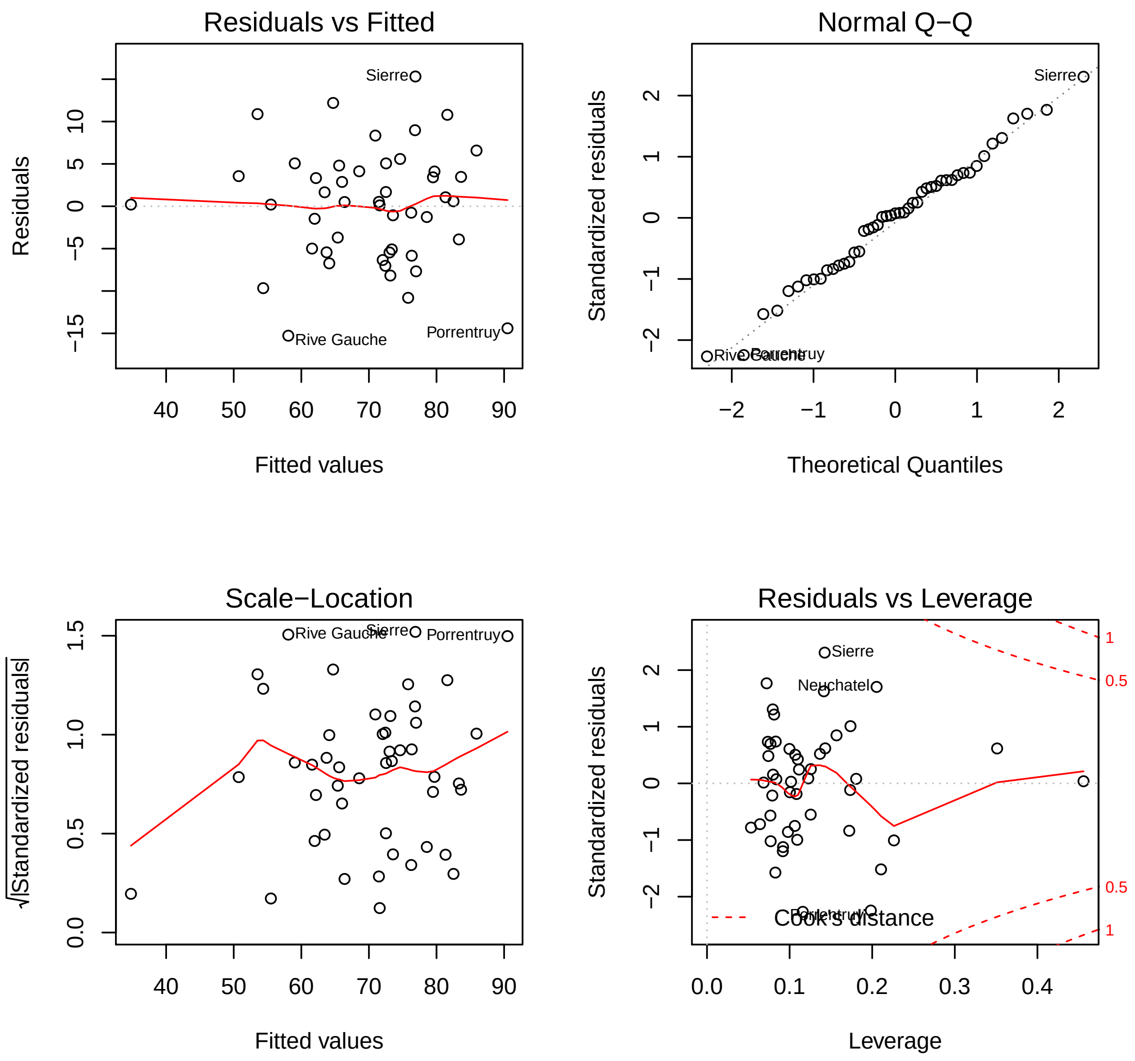
6. In the displayed model, does the variable "Examination" make a significant contribution (at least at the 5% level of significance, use the student table to find the critical value) to the prediction of the dependent variable? (Articulate your response using a formal hypothesis testing)

7. Comment on the model in general: is it statistically satisfactory? You may use the regression output along with the figures. Comment on the over fit, the possible ways to improve the model, and the main hypothesis underlying Ordinary Least Squares.

```
options(width=100)    # limit the width of the table
my.summary <- function(x,...){
  c(mean=mean(x, ...),
    sd=sd(x, ...),
    median=median(x, ...),
    min=min(x, ...),
    max=max(x,...)    ) # ,
    #n=length(x))
}
sapply(swiss, my.summary)

##        Fertility Agriculture Examination Education   Catholic Infant.Mortality
## mean    70.14255    50.65957   16.489362 10.978723   41.14383        19.942553
## sd      12.49170    22.71122    7.977883  9.615407   41.70485         2.912697
## median  70.40000    54.10000   16.000000  8.000000   15.14000        20.000000
## min     35.00000     1.20000    3.000000  1.000000    2.15000        10.800000
## max     92.50000    89.70000   37.000000 53.000000  100.00000        26.600000
```

```
Fertility.lm<-lm(Fertility ~ Agriculture + Examination + Education +
Catholic + Infant.Mortality , data = swiss)
# summary(Fertility.lm)
MySum<-summary(Fertility.lm)
MySum$coef[,1:3]

##                    Estimate  Std. Error   t value
## (Intercept)      66.9151817 10.70603759  6.250229
## Agriculture      -0.1721140  0.07030392 -2.448142
## Examination      -0.2580082  0.25387820 -1.016268
## Education        -0.8709401  0.18302860 -4.758492
## Catholic          0.1041153  0.03525785  2.952969
## Infant.Mortality  1.0770481  0.38171965  2.821568
```

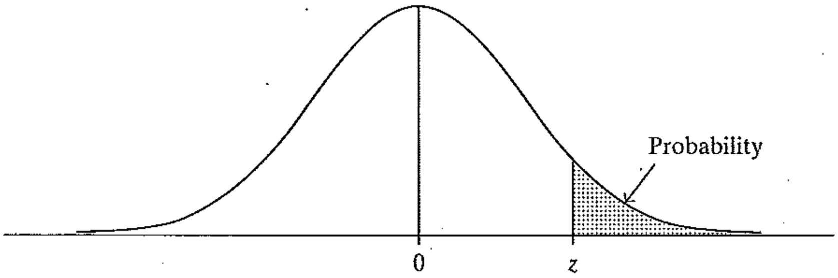
```
MySum$adj.r.squared

## [1] 0.670971

MySum$fstatistic

##    value    numdf    dendf
## 19.76106  5.00000 41.00000

pf(19.76, 5, 41, lower.tail = FALSE)

## [1] 5.598021e-10
```


Residuals vs Fitted
Residuals
Fitted values
−15
−5
0
5
10
40
50
60
70
80
90
Sierre
Rive Gauche
Porrentruy
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
Sierre
Rive Gauche
Porrentruy
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
40
50
60
70
80
90
Rive Gauche
Sierre
Porrentruy
Residuals vs Leverage
Standardized residuals
Leverage
−2
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
Sierre
Neuchatel
Porrentruy
Cook's distance
1
0.5
0.5
1

9 Formula sheet

- $\bar{x} = \frac{\sum_i x_i}{n}$
- $z = \frac{x-\mu}{\sigma}$ *(population)*
- $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$
- $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ *(sample)*
- Confidence intervals:
  - □ $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ *(if normal distribution)*
  - □ $\bar{x} \pm t^{\alpha}_{d.o.f}\frac{s}{\sqrt{n}}$ *(if normal not applicable)*
  - □ $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ *(proportions)*
- Critical values:
  - □ $z = \frac{x\bar{-}\mu}{\frac{\sigma}{\sqrt{n}}}$ *(if normal distribution)*
  - □ $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ *(if normal not applicable)*
  - □ $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ *(proportions)*
- $(x_1 \bar{-} \bar{x}_2) \pm t^{\alpha}_{d.o.f.}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- $\hat{y} = b_0 + b_1 x$
- Probabilities:
  - □ $P(A) + P(A') = 1$
  - □ $P(A \text{ or } B) = P(A \bigcup B) = P(A) + P(B) - P(A \text{ and } B) = P(A) + P(B) - P(A \bigcap B)$
  - □ $P(A|B) = \frac{P(A \text{ and } B}{P(B)} = \frac{P(A \bigcap B}{P(B)}$

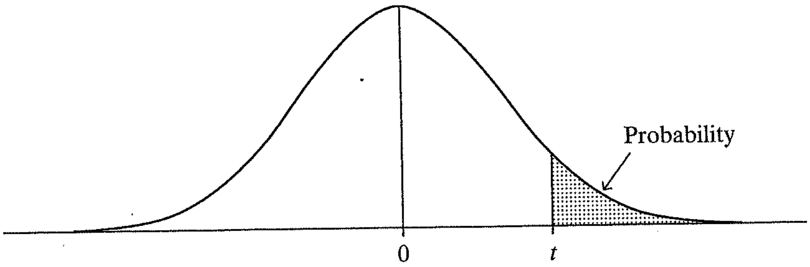
TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z, probabilities are found by symmetry)



	Second Decimal Place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000233									
4.0	.0000317									
4.5	.00000340									
5.0	.000000287									

Source: R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968).

TABLE B: *t* Distribution Critical Values



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.091

Source: "Table of Percentage Points of the *t*-Distribution." Computed by Maxine Merrington, Biometrika, 32 (1941): 300. Reproduced by permission of the Biometrika trustees.