UNIVERSITY OF TORONTO
Faculty of Arts & Science

STA130 F23 Final Examination

Friday, December 19, 2023

Duration: 2 hour and 50 minutes starting 10 minutes after the hour.
Aids Allowed: One "normal" $11 \times 8\frac{1}{2}$ page "cheatsheet", any size, any font, front and back.
Instructions: Write answers in the space provided in the exam and mark multiple choice answers on the scantron sheet at the end of the exam to match the exam question numbers.

Exam Reminders:

- Fill out your name and UTORid on this page and on the scantron answer sheet attached as the last page of the exam.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- As a student, you help create a fair and inclusive writing environment. If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside.

0. To keep things fair, no questions about the exam will be answered during the duration of the exam. If you think there is a problem with a question, note the question and briefly describe the problem in the space below and your concern will be evaluated during marking.

1. For `df` a `pd.DataFrame` with columns `Question_1_Correct` and `Question_2_Correct` containing only `python` values `True` and `False`, the expression `df.Question_1_Correct + df.Question_2_Correct` will result in a `pd.Series` object of type `int`. What is the name of the process that is being carried out to enable this arithmetic operation?

   A. Boolean selection B. Concatenation C. Conditional Logic D. Coercion

2. What are the data types of the values passed into the `dict` in the following `pandas` method call?

```
pd.DataFrame({"column_1":  values_1, "column_2":  values_2})
```

   A. `dict` B. `list` or `tuple` C. `np.array` D. `str`

   E. We can't tell what the data `types` are based on the information provided

3. Suppose `df` is a `pandas DataFrame` object with no missing values that has columns `Country`, `Athlete`, `Event` which has several different repeated values of type `str`, and `Medal` which has only the `python` values `"G"`, `"S"`, and `"B"`. Which of the following `pandas` expressions returns ONLY the number of times the `Event` column is `"400 Meter Sprint"` for each possible value in the `Medal` column?

   A. `df.groupby(["Medal", df.Event=="400 Meter Sprint"]).size()`

   B. `df[df.Event=="400 Meter Sprint"].groupby("Medal").count()`

   C. `(df.groupby("Event")["Medal"]=="400 Meter Sprint").sum()`

   D. `df[["Event", "Medal"]].groupby("Medal").value_counts()`

   E. `df["Medal"][Event=="G"].groupby("Medal").size()`

4. Suppose `df` is a `pandas DataFrame` object with no missing values that has columns `Country`, `Athlete`, `Event` which has several different repeated values of type `str`, and `Medal` which has only the `python` values `"G"`, `"S"`, and `"B"`. Which of the following `pandas` expressions returns the rows of `df` where the countries won only bronze (`"B"`)?

   A. `df[ (df.Medal!="G") & (df.Medal!="S") ]` B. `df[ df.Medal=="B" ]`

   C. `df[ (df.Medal!="G") & (df.Medal!="S") & (df.Medal=="B")) ]`

   D. All of the above E. None of these options

5. Suppose `df` is a `pandas DataFrame` object with no missing values that has columns `Country`, `Athlete`, `Event` which has several different repeated values of type `str`, and `Medal` which has only the `python` values `"G"`, `"S"`, and `"B"`. For the purposes of a `pandas .groupby` operation or a **bar plot**, the information in the `Event` column can be stored as a `Object` or `str` in `pandas`, but from a data analysis perspective what kind of data does the `Event` column contain in a more conceptual sense?

   A. binary, boolean, or logical data B. ordinal qualitative, categorical, or discrete

   C. continuous, numeric, or quantitative D. nominal qualitative, categorical, or discrete

6. With respect to which of the following could all current UofT students be considered (a) a sample versus (b) a population?

   A. (a) A subset of second year statistics majors; (b) Students at large Canadian universities
   B. (a) Students at large Canadian universities; (b) A subset of second year statistics majors
   C. (a) Students at large Canadian universities; (b) Students at public research universities
   D. (a) A Subset of international university students; (b) A subset of second year statistics majors
   E. None of the above as a sample will always be uniformly random

7. Describe the relative benefits of histograms and boxplots, noting drawbacks of each as well; and, additionally describe an alternative data visualization that might be used instead of either of these in the context of a single data sample. *Hint: a pros/cons table might be a help way to organize answers.*

8. Write the definition of a *p-value*.

9. Explain the roles and relationships of *parameters*, *population distributions*, *empirical distributions of samples*, *statistics*, and *sampling distributions* in the context of *statistical inference*.

10. Explain what is meant by the statement, "We have *95% confidence* that the *true parameter* value is contained within the interval (a, b). Why is this not the same as saying, "There's a *95% chance* that the *true parameter* value will be in the interval (a, b)"? After a *confidence interval* is constructed, is it sensible to talk about a *probability* that the *true parameter* is contained within the interval?

11. In what the two ways can we influence the length of a *confidence interval*, and which is more useful?

DO NOT MARK ANSWERS ON YOUR SCANTRON FOR ITEMS #7-11: RESUME MARKING YOUR MULTIPLE CHOICE SCANTRON ANSWERS FROM ITEM #12

12. *Bootstrapping* is used to *simulate* the *sampling variability* (as a *sampling distribution*) of a *statistic*. What is the fundamental assumption that is made to leverage bootstrapping for this purpose?

    A. *Reproducibility* as provided through `np.random.seed(...)`
    B. Using the distribution of the population of the data to understand sampling variability
    C. That the *null hypothesis* (which will depend on the context of the test of interest) is correct
    D. The trick of sampling without replacement which is what leads to the creation of the sampling distribution of the test statistic
    E. Presuming the sample can be used in place of the population and thereby providing a mechanism to simulate samples synthetically in a manner that parallels the sample in question

13. What does `np.random.choice(["a","b","c","d","e","f"], replace=True, size=4)` produce if we instead use `replace=False`?

    A. It will produce a bootstrap sample
    B. It will roll a "fair" six-sided die four times
    C. It will return a shuffled version of original list of data `["a","b","c","d","e","f"]`
    D. It will randomly choose four distinct values from the original list of data
    E. Nothing, the code will not run and will produce an error

14. The *binomial probability mass function* (often referred to as the *binomial distribution*)

$$Pr(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

    is a mathematical expression giving the probability of k "successes" out of n "attempts".

    The $\left(\begin{array}{c} n \\ k \end{array} \right)$ "n choose k" notation counts the unique ways to chose k out of n things based on the order in which the k things sequentially appear, and it is a *combinatorics* calculation that is required so the binomial expression provides the correct probability values.

    The *binomial distribution* is a model of the population of the number of "successes" k there are out of n "attempts", assuming the chance of "success" doesn't change depending on previous or future outcomes. The *binomial distribution* doesn't require any "distributional" assumptions, just the assumption that the performance is consistent. For example, we don't need to say the "shape is normally distributed" or something along those lines to derive the *binomial distribution*.

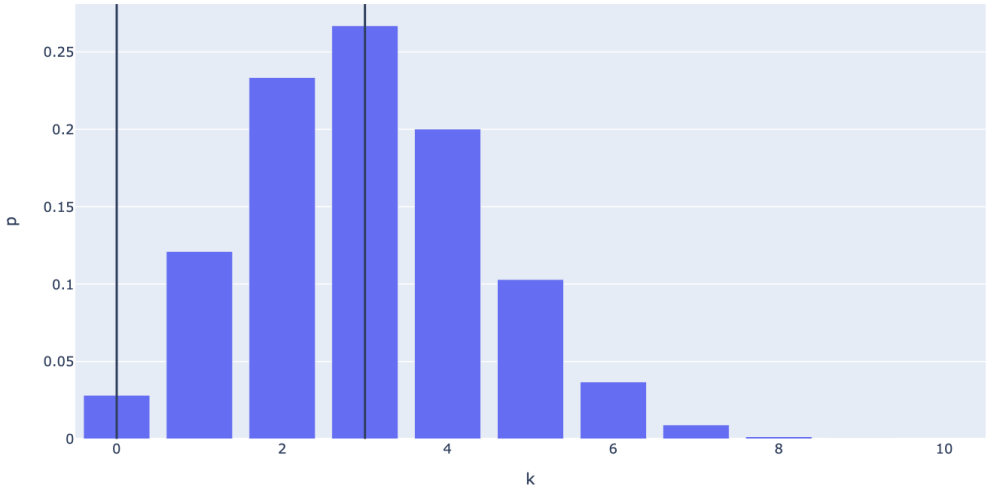
    We could thus model your (assumed "pretty good") ability to "not *miss* basketball shots" with the

    null hypothesis H_0 : the chance you *miss* a shot is 30% (so $p = 0.3$ for the *binomial model*)

    The probabilities of the various number of k shots you would *miss* out of $n = 10$ tries (possibly 0 up to 10) under the consistent performance assumption of (*miss* is a "success") *binomial model* then are

```
>>> from scipy import stats; import numpy as np
>>> stats.binom(p=0.3,n=10).pmf(np.linspace(0,10,11,dtype=int))
array([0.0282475249, 0.1210608210, 0.233474441,
       0.2668279320, 0.2001209490, 0.102919345, 0.036756909,
       0.0090016920, 0.0014467005, 0.000137781, 0.000005905]) # these sum to 1
```

The following is a *barplot* of this *binomial distribution* with vertical lines indicating the observed *test statistic* and expected number of *misses* (out of 10) according to H_0. The next questions ask about which probabilities should be added together to calculate the theoretical *p-value* for this observed test statistic (of $k = 0$ *missed* shots out of $n = 10$ attempts).



Using the characterization given in the table below, what is the strength of evidence against H_0 for the previous question if "as or more extreme" is "two-sided" so that *p-value* calculations include performances that are both better than and worse than the performance expected by H_0?

	p-value	Evidence against the Null Hypothesis
A.	$p > 0.10$	No Evidence
B.	$0.05 < p < 0.10$	Weak Evidence
C.	$0.01 < p < 0.05$	Moderate Evidence
D.	$0.001 < p < 0.01$	Strong Evidence
E.	$p < 0.001$	Very Strong Evidence

A. None B. Weak C. Moderate D. Strong E. Very Strong

15. What would the strength evidence against H_0 be for the previous question if "as or more extreme" is "one-sided" and only refers to performances with as few or fewer *misses* than what was observed?

A. None B. Weak C. Moderate D. Strong E. Very Strong

16. Which of the following is the best statement about evidence against the *null hypothesis* H_0 for the previous problems where the observed test statistic is 0 *missed* shots out of 10 attempts?

    A. It is extremely strong since it shows "perfect accuracy" and thus *sufficiently justifies* the conclusion "you shoot with perfect accuracy"

    B. It is extremely strong since it shows "perfect accuracy" and thus *proves* the conclusion "you shoot with perfect accuracy"

    C. The *p-value* calculation quantifies the degree of evidence *in favor of* the conclusion "you shoot with perfect accuracy"

    D. The strength of evidence it provides *against* H_0 depends on the denominator of $n = 10$

    E. It is meaningless because the "perfect accuracy" assumption that the chance of missing is $p = 0$ is not possible

17. Which of the following is a true statement with respect to *sample size* and the *simulation size* used to create an estimated *sampling distribution*?

    A. Increasing *sample size* provides more accurate estimation of a *sampling distribution* and *p-value*

    B. Increasing both the *sample size* and the *simulation size* will make a *theoretical p-value* computation (like from the *binomial distribution* considered in the previous problems) more accurate

    C. Increasing the *simulation size* increasingly reduces *statistical inference estimation uncertainty*

    D. The *variability* of the *theoretical sampling distribution* is a function of the *sample size* whereas the accuracy of the *estimation* of the *sampling distribution* is a function of the *simulation size*

    E. *Sample size* is easier to change than the *simulation size* once data is collected and being analyzed

18. Provide `python` pseudocode to *simulate* a *p-value* which *approximates* the "two sided" *theoretical p-value* for the *binomial distribution* considered in the previous few problems for an *observed test statistic* of 0 misses, and comment on how the simulation size influences the *approximation accuracy*.

19. Give `python` psuedocode specifying reasonable *population distribution models* for each of the following

    - Heights of students at UofT
    - The number of correct answers on a quiz with 10 questions
    - An "unfair" six-sided die with sides that don't have equal probability of being rolled

    and then briefly describe a general *parametric* method that could be used test the hypothesized *mean parameters* of these specifications, and a general *nonparametric* method that could be used to provide *statistical inference (estimation)* of the *parameters* of these populations.

20. For contexts in which we are interested in *population means*, and we have samples `x1` and `x2` each stored as an `np.array`, write the *null hypotheses* and corresponding *observed test statistic* calculations for *hypothesis testing* involving *one sample*, *two sample*, and *paired sample* contexts.

    Hint: the respective sampling distributions corresponding to each of these could be simulated using synthetic samples produced under assumptions of the null hypothesis, repeated permutation of group labels, and repeated random assignment of the group membership of the sampled pairs.

21. Suppose you have a sample of size $n = 2$ numbers $x_{(1)} < x_{(2)}$. Assuming sample quantiles defined as

$$\texttt{quantiles}(x_{(i)}) = \frac{\left(\frac{1}{n}\sum_{j=1}^{n} 1_{[x_{(j)}<x_{(i)}]}(x_{(i)})\right) + \left(\frac{1}{n}\sum_{j=1}^{n} 1_{[x_{(j)}\leq x_{(i)}]}(x_{(i)})\right)}{2}$$

draw a *boxplot* of this sample and explicitly note the rationale for your chosen treatment of *whiskers* and *outliers* in your visualization. Then draw the *bootstrap sampling distribution* of the *median* of this sample assuming an extremely large number of bootstrap samples (on a relative proportional scale as opposed to an absolute count scale), and annotate this figure with visual lines indicating the *exactly precise* numerical locations of the end points of the *50% confidence interval* (which are the values that 50% of the *bootstrapped sampling distribution* are equal to or between). Report the length of this *50% confidence interval*, and label your x and y axes and their numeric scales for both figures.

22. Explain what *survivor bias* means for UofT students by discussing its possible role in differences in study habits that might exist between a sample of first year students currently enrolled at UofT as opposed to a sample of fourth year students currently enrolled at UofT.

DO NOT MARK ANSWERS ON YOUR SCANTRON FOR ITEMS #18-22: RESUME MARKING YOUR MULTIPLE CHOICE SCANTRON ANSWERS FROM ITEM #23

23. From "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine" published in the New England Journal of Medicine in 2020 by Polack et al. of the C4591001 Clinical Trial Group:

> A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of [laboratory-confirmed] Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo. ... Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient.

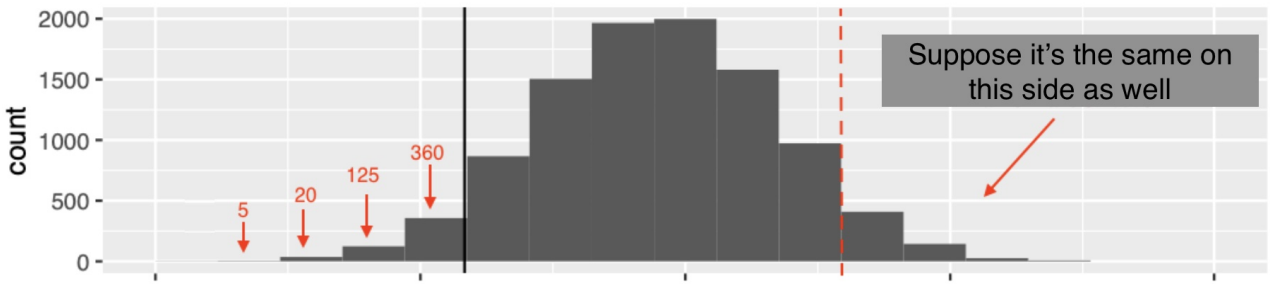
This gives a vaccine efficacy of $\tilde{x} = 1 - (8/21720)/(162/21728) = 1 - p_t/p_c \approx 95\%$ with an associated *p-value* that was less than 0.025. Which of the following best describes the *test statistic* $\tilde{x} \approx 95\%$?

A. The probability that the BNT162b2 injection stops a Covid-19 infection
B. The ratio of the absolute difference between the probabilities of laboratory-confirmed Covid-19 in the placebo and BNT162b2 groups divided by the maximum of the two probabilities
C. The relative proportional ratio of the rate of non-severe laboratory-confirmed Covid-19 cases in the BNT162b2 treatment group over the analogous rate for the placebo control group
D. The chance a Covid-19 infection in the study occurred in the placebo group
E. The difference between the probability a Covid-19 infection in the BNT162b2 and placebo groups

24. Following up on the previous question, which of the following *alternative hypotheses* would most appropriately correspond to an analysis with no prior belief about whether or not there was a difference between the chance of laboratory-confirmed Covid-19 cases in the BNT162b2 and placebo groups?

A. $H_A : p_t = p_c$ B. $H_A : p_t > p_c$ C. $H_A : p_t < p_c$ D. $H_A : H_0$ is false

25. Following up on the previous questions, suppose the *p-value* for this study was based on the following *simulated sampling distribution* of the *test statistic* with $10,000$ *simulated test statistics*. Assuming the *p-value* is based on a *symmetric* view of "as or more extreme", what is the "tallest" bin the *test statistic* could have fallen into while still definitely producing the result of this study?



A. The "5" bin B. The "20" bin C. The "125" bin D. The "360" bin

26. Following up on the previous questions, what was the chance of getting laboratory-confirmed Covid-19 if you were a participant in placebo group of this study?

A. 8/21720 B. 162/21728 C. 151/162 D. 1/8 E. 10/43448

27. Following up on the previous questions, what was the chance of getting severe Covid-19 if you were a participant in this study?

A. 8/21720 B. 162/21728 C. 151/162 D. 1/8 E. 10/43448

28. Following up on the previous questions, given that someone was a laboratory-confirmed Covid-19 case in this study, the chances of being a severe case were $p_{s|t} = 1/8 = 0.125 > p_{s|c} = 9/162 = 0.0555...$, or about 2.25 times higher for the BNT162b2 group versus the placebo group. Which of the following is an appropriate conclusion about this data and its natural *test statistic*?

A. The *test statistic* $\frac{n_{s|t}}{n_t} - \frac{n_{s|c}}{n_c} = 1/8 - 9/162$ is too complicated so there's no way to derive its *sampling distribution* and hence no way to use it as evidence in a *hypothesis testing* context

B. The $\frac{n_{s|t}}{n_t} - \frac{n_{s|c}}{n_c}$ depends on very rare events and therefore may suffer from using insufficient data

C. The $\frac{n_{s|t}}{n_t} - \frac{n_{s|c}}{n_c}$ *test statistic* should be dismissed because contradicts the *efficacy test statistic* $\tilde{x} = 1 - (8/21720)/(162/21728) = 1 - p_t/p_c$ already giving sufficient evidence of vaccine effectiveness

D. The $\frac{n_{s|t}}{n_t} - \frac{n_{s|c}}{n_c}$ *test statistic* should be ignored since it doesn't give evidence in favor of vaccines

E. This is statistical evidence contradicting the claim that vaccines are effective against Covid-19

29. Following up on the previous questions, with respect to the differential nature of *relative risk* versus *absolute risk*, which of the following address the potential implications of the *generalizability* of *relative risk* in conjunction with a lack of *generalizability* of *absolute risk* for BNT162b2 vaccines?

A. *Relative risk* has increasingly widespread impacts as *absolute risk* (a primary concern) increases

B. The study gives evidence of *relative risk* reduction, but some of the data suggest vaccines may also carry some risk; so, vaccination should be a choice based on personal *absolute risk* tolerance

C. The study gives evidence of *relative risk* reduction, so even if some people choose not to vaccinate, this is not a concern for those who vaccinate since they'll have almost no *absolute risk*

D. The study gives convincing evidence of *relative risk* reduction, which is exactly why vaccination must be mandated for everyone since this will make *absolute risk* as low as possible for everyone

E. This study allows us to accurately predict *absolute risk* on the basis of its *relative risk* estimates

30. Over the last few years there have been a notable a number of protests in France. Somewhat recently, protests in response to a proposed increase in the age at which workers become eligible for retirement benefits received considerable media attention; before that there had been protests in relation to some political unrest over president Macron's government's energy policies in response to economic shortages associated with the conflict in Ukraine; and before that there had been protests in response to Covid-19 related mandates. There seemed to have been so many protests in France that it became pretty easy to come across internet jokes along the lines of, "Learn how to protest from the French".

Information about the nature of protest incidents, organizations involved, and references to news outlet reports on the protests is available from acleddata.com. In particular, the location and sizes

of the protests are available from this data source. Protest sizes are stored in a "tags" column that usually has a legible note about "crowd size". This is often a numerical value, but sometimes the numeric values are written as words (like "twenty") or rough approximations or ranges. By translating words into their numeric values and/or taking an average of the range limits, nearly all of the protest events can be given a single numerical crowd size value, and this should hopefully give a fairly accurate transformation of the information in the "crowd size" tag into numeric values. So, with this data in hand let's consider the relationship between location and protest size in France.

Below is an analysis that uses the same linear structure as a statistical methodology known as ANOVA (shorthand for ANalysis Of VAriance). Statistical tests related to ANOVA can be derived as a book keeping process that partitions the variation in the data in an explanatory manner; however, the same predictive model can also be viewed as a special case of multiple linear regression.

```
# keep only data from city with >=300 protests
locations_w300ormore_protests = france_protests.location.\
    apply(lambda x: x in locations[locations>299].index)
france_protests_locations_sizes_nonan = \
    france_protests.loc[locations_w300ormore_protests,
                        ['location','tags2float']].dropna()
france_protests_locations_sizes_nonan
```

	location	tags2float
0	Paris	30.0
3	Paris	200.0
41	Lyon	200.0
71	Toulouse	100.0
75	Bordeaux	150.0
...	...	...
25256	Rennes	300.0
25259	Toulouse	100.0
25260	Marseille	4000.0
25268	Lyon	300.0
25283	Paris	100.0

2621 rows × 2 columns

```
import pandas as pd
# Paris indicator not included: this makes it the "baseline"
# Thus the value of the intercept when all other city indicators
# are "off" reflects Paris
yX = pd.get_dummies(france_protests_locations_sizes_nonan,
                    drop_first=False, prefix='', prefix_sep='').\
    drop('Paris', axis=1)
yX
```

	tags2float	Bordeaux	Lyon	Marseille	Montpellier	Nantes	Rennes	Toulouse
0	30.0	0	0	0	0	0	0	0
3	200.0	0	0	0	0	0	0	0
41	200.0	0	1	0	0	0	0	0
71	100.0	0	0	0	0	0	0	1
75	150.0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
25256	300.0	0	0	0	0	0	1	0
25259	100.0	0	0	0	0	0	0	1
25260	4000.0	0	0	1	0	0	0	0
25268	300.0	0	1	0	0	0	0	0
25283	100.0	0	0	0	0	0	0	0

2621 rows × 8 columns

```
locations = france_protests.location.value_counts()
locations[locations>299]
```

```
Paris          1175
Toulouse        468
Marseille       428
Lyon            422
Bordeaux        403
Nantes          363
Rennes          334
Montpellier     313
Name: location, dtype: int64
```

```
import numpy as np
X_ = pd.concat([0*yX.iloc[:,1:],
                pd.DataFrame(np.eye(7),columns=yX.columns[1:])],
               ignore_index=True)
X_
```

	Bordeaux	Lyon	Marseille	Montpellier	Nantes	Rennes	Toulouse
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	1.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	1.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	1.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	1.0

```
from sklearn.linear_model import LinearRegression
y = yX.iloc[:,0]
X = yX.iloc[:,1:]
reg = LinearRegression(fit_intercept=True).fit(X, y)
reg.intercept_, reg.coef_ # all sklearn gives is y-hats
```

```
(8909.189107413016,
 array([-6457.96123856, -6941.43989111, -2347.59854138, -7043.93695224,
        -6552.13313726, -7276.40281709, -5072.20768326]))
```

```
reg.predict(X_)
```

```
array([8909.18910741, 2451.22786885, 1967.7492163 , 6561.59056604,
       1865.25215517, 2357.05597015, 1632.78629032, 3836.98142415])
```

Which of these cities is suggested to have the largest protests according to the fitted coefficients of the (multiple linear regression) ANOVA specification model (and subsequent predictions) given above?

A. Paris B. Bordeaux C. Lyon D. Marseille E. Montpellier

31. According to this model fit, how many more protesters are there on average in a protest in Marseille compared to a protest in Bordeaux (as a rounded integer number)?

A. 4110 B. 6941 C. -6941 D. 6561

32. How could we go about giving statistical evidence of differences in protest sizes between cities using this model specification? Describe the next steps you would take to follow up on the analysis above and which comparisons you could address statistically with this current model specification.

33. Give a possible *confounder* variable (that isn't included in the model but) that might go some way in explaining the difference in the average number of protesters in the different cities.

34. The statistical model currently under consideration can be specified as

$$y_i = \overbrace{\beta_0 + \left[\sum_{\substack{c \text{ in} \\ \text{cities} \\ \text{except} \\ \text{Paris}}} \beta_c 1_{[c]}(\text{city}_i)\right]}^{E[y_i|\text{city}_i]} + \epsilon_i$$

Write the *null hypotheses* for statistically analyzing each of the *coefficients* of this model specification.

35. What would it mean in practical terms if all these *null hypotheses* were together simultaneously true? And how about if we asked this same question but without including the *null hypotheses* for β_0?

DO NOT MARK ANSWERS ON YOUR SCANTRON FOR ITEMS #32-35: RESUME MARKING YOUR MULTIPLE CHOICE SCANTRON ANSWERS FROM ITEM #36

36. The histograms below shows the model residuals $\hat{\epsilon} = y - \hat{y}$ for fitted values from the fit of this model

$$\hat{y} = \hat{\beta}_0 + \sum_{c \text{ in cities}} \hat{\beta}_c 1_{[c]}(\text{city})$$

An assumption of the statistical analysis of linear regression (and ANOVA) is that the population of the error terms ϵ_i (which the residuals $\hat{\epsilon}_i$ estimate) is normally distributed. The histograms of the residuals below suggest that the normality assumption of the error terms is inappropriate due to...

A. strong left skew and extreme outliers
B. heteroskedasticity
C. strong right skew and extreme outliers
D. None of these: the assumption is appropriate

```
from plotly.subplots import make_subplots
import plotly.graph_objects as go
fig = make_subplots(rows=1, cols=2, row_heights=[0.5],
                    subplot_titles=('residuals', 'zoomed in residuals (left side)'))
fig.update_layout(height=350, width=750)
fig.add_trace(go.Histogram(x = y-reg.predict(X)), row=1, col=1)
fig.add_trace(go.Histogram(x = y-reg.predict(X)), row=1, col=2)
fig.update_xaxes(range=[-20000, 20000], row=1, col=2)
fig.show()
```

37. Which of the following best characterizes the appropriateness of the normality assumption relative to the last question when the log of the size of the protest is the outcome as shown below? It is...

A. about the same
B. somewhat worse
C. somewhat improved
D. quite improved but still questionable

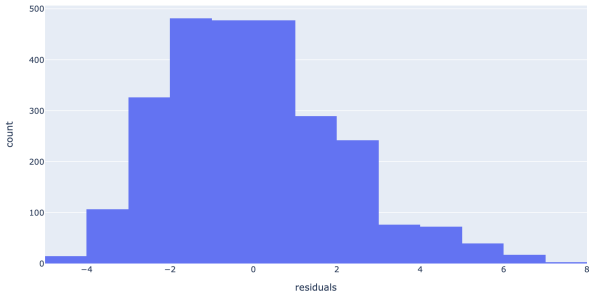
```
log_y = yX.iloc[:,0].apply(np.log)
reg = LinearRegression(fit_intercept=True).fit(X, log_y)
reg.intercept_, reg.coef_ # all sklearn gives is y-hats
```

```
(5.775348815291884,
 array([-0.17081997, -0.0200198 ,  0.17085054, -0.16558324,  0.01268541,
        -0.11511613, -0.33082615]))
```

```
X.columns
```

```
Index(['location_Bordeaux', 'location_Lyon', 'location_Marseille',
       'location_Montpellier', 'location_Nantes', 'location_Rennes',
       'location_Toulouse'],
      dtype='object')
```

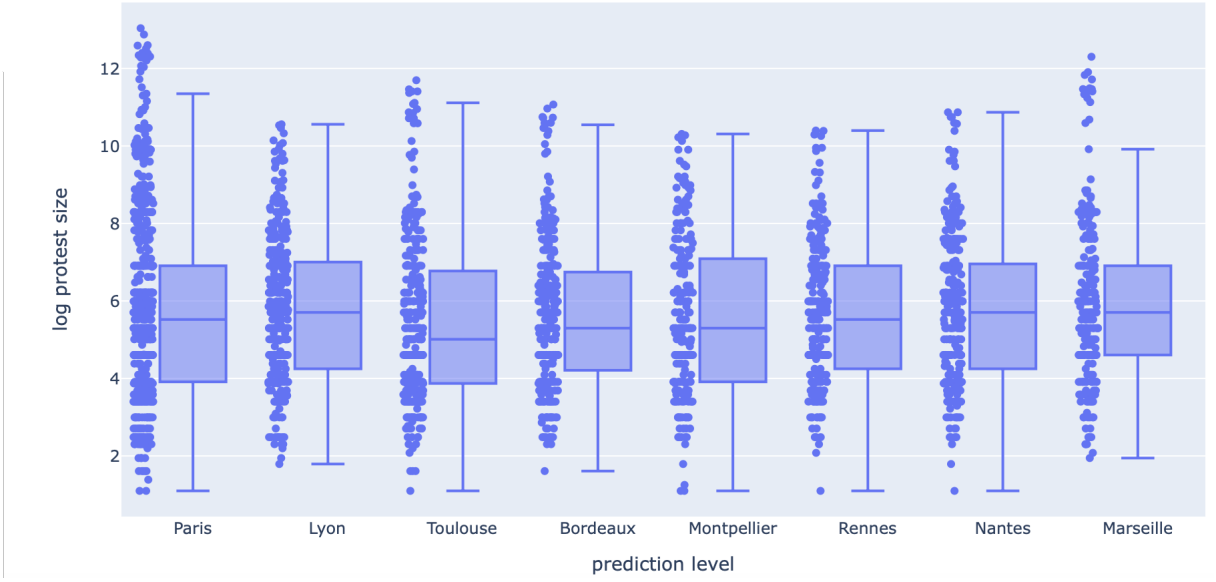
```
import plotly.express as px
fig = px.histogram(pd.DataFrame({"residuals": log_y-reg.predict(X)}), x="residuals", nbins=20)
fig
```



38. Another assumption of the statistical analysis of linear regression (and ANOVA) is that the error terms ϵ_i (which the residuals $\hat{\epsilon}_i$ estimate) are *homoskedastic* so there is not a difference in the variability (the variance) of the error terms across the different prediction levels. The boxplots and swarm plots below show the distributions of the log protest size across every prediction level. What does this suggest about this *heteroskedasticity* assumption for the updated model? That it is...

A. unreasonable
B. not perfect but fairly reasonable
C. likely plausible
D. Almost certainly plausible

```
df = pd.DataFrame({"log protest size": log_y,
                   "prediction level": france_protests_locations_sizes_nonan.location})
fig = px.box(df, x="prediction level", y="log protest size", points="all")
fig.show()
```



39. The statistical assumptions for a statistical analysis of our linear regression (ANOVA) model

$$y_i = \overbrace{\beta_0 + \left[\sum_{\substack{c \text{ in} \\ \text{cities} \\ \text{except} \\ \text{Paris}}} \beta_c 1_{[c]}(\text{city}_i)\right]}^{E[y_i|\text{city}_i]} + \epsilon_i \qquad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma)$$

are that the city of a protest is known (correctly), and the previously considered assumptions of *normality* and *homoskedasticity* of the error terms, which are expressed mathematically with $\mathcal{N}$ indicating normality and the *zero mean* and constant *standard deviation* σ implying *homoskedasticity*.

There is in fact one additional assumption that is necessary for a straight forward statistical analysis of this model, which is reflected in the *i.i.d.* notation above which stands for *independently and identically distributed*. This is the assumption that the variability away from the expected value of every observation does not depend on the actualized values of any of the other observations.

Do you think the *i.i.d.* assumption is a reasonable assumption to apply to our protest data here?

A. Yes, in statistics we always make the assumption that observations in a sample are independent
B. Yes, because it's hard to assume the size of one protest could affect the size of another protest
C. It would be unreasonable if the protests were driven by underlying social dynamics, such as a growing social movement with each protest attracting new participants or new counter-protests
D. There's no way to really know this, so this is not a consideration that needs to be entertained

40. Is the *p-value* that would be used to evaluate the *null hypotheses* regarding the coefficients in the linear regression (ANOVA) model context *theoretical* or *simulated*? And is it *parametric* or *nonparametric*?

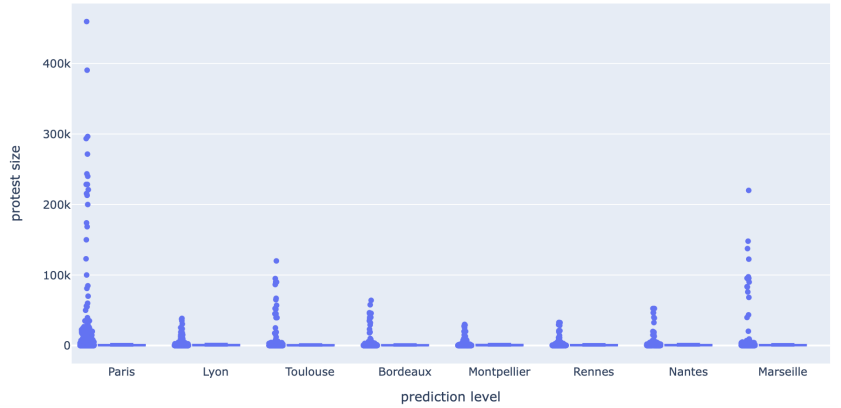
A. theoretical parametric B. simulated parametric
C. theoretical nonparametric D. simulated nonparametric E. none of the previous options

41. Returning to the two outcomes considered so far of using either the size of the protest or the log of the size of the protest, do the corresponding point estimates of the coefficients appear to agree?

A. Yes, both model predictions are equivalent with one being only a log transformation of the other
B. Yes, the rank ordering of cities by average protest size predicted by both models is the same
C. No, for example protest size in Marseille and Nantes relative to Paris changes between models
D. We don't have enough information to answer this on the basis of the provided model printouts

42. Use the plot here compared to the previous plot to explain the conclusion of the previous problem.

```
df = pd.DataFrame({"protest size": y,
                   "prediction level": france_protests_locations_sizes_nonan.location})
fig = px.box(df, x="prediction level", y="protest size", points="all")
fig.show()
```



43. Given the considerations of the last few problems, a statistical analysis of this data with the model specification being considered may be more or less appropriate depending on the correctness of the assumptions entailed in using this model. Regardless, an evaluation of the reliability of any conclusions we might draw from the data is certainly a reasonable and sensible pursuit. And, the issues of reliability and generalizability are even more relevant if we expand the number of cities under consideration. So far we've only considered the eight cities with the most data, but if we were to expand our analysis to include cities for which we have data on at least 50 protests then there would be 110 cities under examination. Why is the reliability and generalizability of an analysis based on 110 cities of greater concern than an analysis based on 8 cities?

```
1 locations = france_riots.location.value_counts()
2 locations[locations>49]
```

```
Paris           1175
Toulouse         468
Marseille        428
Lyon             422
Bordeaux         403
                ...
Saint-Malo        54
Valenciennes      54
Guingamp          54
Arras             52
Pontivy           51
Name: location, Length: 110, dtype: int64
```

44. Describe a *machine learning* approach to determining the number of French cities that we would be willing to predict average protest size for using this data, and discuss (for this data in particular) the relative drawbacks of this compared to drawbacks of a statistical hypothesis testing based approach.

DO NOT MARK ANSWERS ON YOUR SCANTRON FOR ITEMS #42-44: RESUME MARKING YOUR MULTIPLE CHOICE SCANTRON ANSWERS FROM ITEM #45

45. Suppose you have a three-level categorical variable x taking on the values "a", "b", and "c" and you're going to use this variable to predict continuous outcome y. Which of the following added to $E[y] = \beta_0 + \cdots$ uniquely completes the mathematical form of the specification here?

    Hint: by "uniquely" we mean that a is defined "uniquely" for $a = 1$ whereas for $a + b = 1$ it is not.

    A. $\beta_x x$ B. $\beta_a 1_{[\text{“}a\text{”}]}(x) + \beta_b 1_{[\text{“}b\text{”}]}(x)$ C. $\beta_a 1_{[\text{“}a\text{”}]}(x) + \beta_b 1_{[\text{“}b\text{”}]}(x) + \beta_a 1_{[\text{“}c\text{”}]}(x)$

    D. $\beta_b 1_{[\text{“}b\text{”}]}(x) + \beta_c 1_{[\text{“}c\text{”}]}(x)$ E. Either option B or option D

46. Following up on the same setting of the previous problem, suppose we also had a continuous variable z. What kind of model would the "linear formula" specification `y ~ C(x)*z` define?

    A. It would model parallel linear relationship between z and y that could shift vertically for groups
    B. It would model linear relationship between z and y uniquely within each of the groups
    C. It would model two linear relationships each possibly having unique intercepts and slopes
    D. It would make three possibly differently slopped lines with the same intercept
    E. It would make three unique predictions of the form β_0, $\beta_0 + \beta_1$, and $\beta_0 + \beta_2$

47. Following up on the same setting of the previous problems, assuming all *p-values* were only "very large" or "very small", which of the following results for a `y ~ C(x)*z` specification would suggest that a model of two parallel lines might be appropriate?

    A. Very large *p-values* only for coefficients of two *indicator variables* and two *interaction variables*
    B. Very small *p-values* only for coefficients of two *indicator variables* and two *interaction variables*
    C. Very large *p-values* only for coefficients of one *indicator variable* and one *interaction variable*
    D. Very small *p-values* only for coefficients of one *indicator variable* and one *interaction variable*
    E. None of the above

48. Following up on the same setting of the previous problems, which of the following sets of columns should be added to a column of "ones" to create a *design matrix* X specifying the "two parallel lines" model of the last problem?

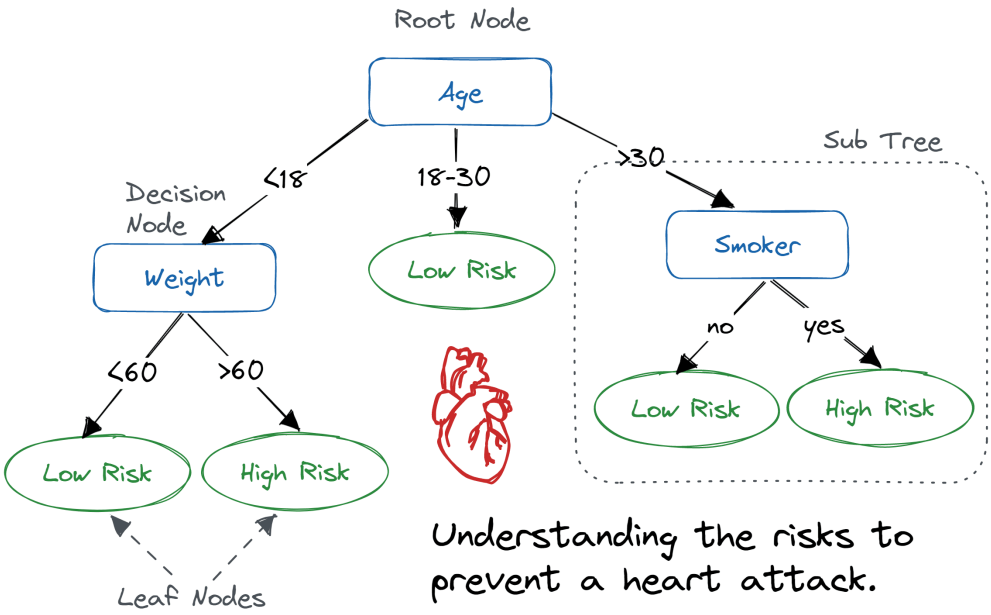
    A. z and $1_{[\text{“}a\text{”}]}(x)$ B. z and $z \times 1_{[\text{“}a\text{”}]}(x)$ C. $1_{[\text{“}a\text{”}]}(x)$ and $z \times 1_{[\text{“}a\text{”}]}(x)$
    D. z and $1_{[\text{“}a\text{”}]}(x)$ and $z \times 1_{[\text{“}a\text{”}]}(x)$ E. None of these options

49. Which of following gives the most general distinction between machine learning (ML) and statistics?

    A. The focus on classification of categorical as opposed to regression of continuous outcomes
    B. The focus on decision tree as opposed to multiple linear regression model predictions
    C. The focus on performance evaluation as opposed to evidence and uncertainty characterization
    D. The focus on binary variables as opposed to indicator and categorical variables
    E. None of the above as differences between ML and stats is more subtle than these options

50. The decision tree below has depth three, but it also initially makes three splits (rather than two) based on the *Age* variable. If the we only allowed each decision node to only split the data into two subsets of data (as opposed to three), how deep would this decision tree be?

A. 2 B. 3 C. 4 D. 5



51. Following up on the previous question, if "High Risk" is a *positive*, which of these is a *false positive*?

A. A non-smoker with weight 70 and age 21 B. A smoker with weight 70 and age 21
C. A non-smoker with weight 50 and age 17 D. A smoker with weight 70 and age 17
E. There's not enough information to tell

52. Following up on the previous questions, what prediction does the decision tree under consideration make for a non-smoker with weight 70 and age 21?

A. Low Risk B. High Risk C. False Positive D. True Negative E. None of these options

53. *Machine learning algorithms* have *tuning parameters* that control *model complexity* and hence subsequently determine the flexibility of *model predictions*. Which of the following could be used to define "stopping rules" to control the *complexity* of the *decision tree model* above if we were constructing the *tree* with more *features* that we thought might be helpful to improve the accuracy of the model?

A. Leaf node sizes prior to a proposed split B. Leaf node sizes that would result from a split
C. The depth of the tree D. The total number of nodes in the tree E. All of these options

54. Explain why a *false negative* is a worse mistake than a *false positive* in predicting heart attack risk.

55. Give an example of a situation where a *false positive* is worse than a *false negative*.

56. Draw a graph with "Age" on the x-axis and "Weight" on the y-axis, and partition the space according to the decision tree of the previous problem, labeling the regions of "High Risk" and "Low Risk" or annotating the figure to indicate the role of smoking in this distinction where appropriate.

57. Draw a *confusion matrix*, labelling *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FP) such that the total number of data points is $n = 100$ and the *sensitivity* ($\frac{TP}{TP+FN}$) is 80%, the *specificity* ($\frac{TN}{NP+FP}$) is 80%, and the *precision* ($\frac{TP}{TP+FP}$) is 80%.

58. Following up on the previous question, this *confusion matrix* will have been the result of an *algorithm* that in fact *predicts* the *probability* of a *positive* event, with the *positive classification predictions* being situations where the predicted probability has exceeded a chosen *threshold* (such as if the predicted probability is greater than 50%). This *threshold* is something that may be changed to influence the number of FP and FN predictions. Describe how we should change the *threshold* if we were more concerned about *false negative* (FN) errors, so that having a higher *sensitivity* was desirable. In answering, give examples of how the the *confusion matrix* above might change in a way that results in maintaining, decreasing, and increasing *precision*, and comment on the likely effect of the proposed *threshold* change on *specificity*.

59. Consider the following recidivism classification matrices (of convicted criminals tendency to reoffend):

Social group A	Recidivates	Does NOT recidivate
Predicted to recidivate	59	13
Predicted to NOT recidivate	13	15

Social group B	Recidivates	Does NOT recidivate
Predicted to recidivate	17	13
Predicted to NOT recidivate	13	57

What is an argument against using *demographic parity*, making an equal proportion of positive predictions across both social groups (A and B), as a measure for fairness of a predictive algorithm?

DO NOT MARK ANSWERS ON YOUR SCANTRON FOR ITEMS #54-59: RESUME MARKING YOUR MULTIPLE CHOICE SCANTRON ANSWERS FROM ITEM #60

60. Which of the following is true of the algorithm giving the recidivism classification matrices above?

A. The algorithm is *equally calibrated*: for both groups the chance a positive prediction is correct equals the true chance the observation is positive, and this is analogously true for negatives, too

B. This algorithm satisfies the *equalized-odds* fairness criterion: both groups have equal proportions of false positive and false negative predictions

C. Both (A) and (B) D. Neither (A) nor (B)

61. If algorithm X satisfies both the fairness criteria of *equalized odds* and is equally *calibrated* with respect to both social groups A and B, which of the following MUST be true?

A. The algorithm makes no prediction errors, so it is *equally calibrated* and satisfies *equalized odds*

B. The base rates (the true chances of positive and negative outcomes) must be correspondingly the same in group A and group B, because if this is so then having the same number of false positives and false negatives (i.e., *equalized odds*) will then additionally imply *equal calibration*

C. Both (A) and (B) must be true D. None of these options