# EKT-812 Problem Set 1: Suggested Solutions

*There are 160 points available on this problem set, and it will be graded out of 120. That is, there are 40 bonus points available.*

**Due Date: Wednesday, February 20**

## Probability Review

1. Consider the following joint distribution:

| ↓ X Y→ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 0 | 0.1 | 0.05 | 0.025 | 0.025 |
| 1 | 0.07 | 0.13 | 0.04 | 0.06 |
| 2 | 0.1 | 0.1 | 0.25 | 0.05 |

(a) Find the first two moments - the mean and variance - of the marginal distributions of $X$ and $Y$. Check that the identity $V[X] = E[X^2] - E[X]^2$ holds.

**Answer:** We have $E[X] = 1.3$ and $E[Y] = 2.315$. We also have $E[X^2] = 2.3$, so $V[X] = 2.3 - 1.3^2 = 0.61$.

(b) Confirm that $E[XY] = 3.19$. Use this in combination with your calculations in (a) above to find $\text{cov}(X, Y)$.

**Answer:** We have $\text{cov}(X, Y) = 3.19 - 1.3 \times 2.315 = 0.1805$.

(c) Find the conditional probability mass function $P(Y = y|X = x)$ and use it to calculate the conditional mean function $E[Y|X = x]$ for each possible realization of $X$.

**Answer:** The conditional distribution of $Y|X$ is as follows

| | $Y = 1$ | $Y = 2$ | $Y = 3$ | $Y = 4$ |
|---|---|---|---|---|
| $X = 0$ | 0.5 | 0.25 | 0.125 | 0.125 |
| $X = 1$ | 0.233 | 0.433 | 0.133 | 0.2 |
| $X = 2$ | 0.2 | 0.2 | 0.5 | 0.1 |

So, we have

- $E[Y|X = 0] = 1.875$
- $E[Y|X = 1] = 2.3$
- $E[Y|X = 2] = 2.5$

(d) Calculate $E[E[Y|X = x]]$. With respect to which distribution is the outer expectation taken? The inner expectation? What do you notice about your answer as compared to the means you calculated in (a) above?

**Answer:** The outer expectation is taken with respect to the marginal distribution of $X$. The inner expectation is with respect to the conditional distribution $P(Y = y|X = x)$. You should notice that $E[E[Y|X = x]] = E[Y] = 2.315$; this is a consequence of the law of iterated expectations.

[3 + 2 + 3 + 2 = 10 pts]

2. Consider a population of workers, each with two possible skills: call them $s_0$ and $s_1$. Think of $s_0$ as, say, programming skill and $s_1$ as language skill, which is useful for becoming e.g. a lawyer. Assume that earnings are determined as follows: there is a prevailing wage $w_0 > 0$ in the tech sector and $w_1 > 0$ in

the legal sector. Workers are paid by skill, so more skilled workers earn more, and in particular,

$$Y_0 = w_0 e^{s_0} \quad (1)$$
$$Y_1 = w_1 e^{s_1} \quad (2)$$

where $Y_0$ and $Y_1$ are *potential* earnings in each sector.

Assume that the joint distribution of skills in the population is bivariate normal, with $E[s_0] = E[s_1] = 0$. (Here skills are on a logarithmic scale, so the fact that they can be negative is not a problem.) Let the standard deviation of $s_0$ be $\sigma_0$, the standard deviation of $s_1$ be $\sigma_1$, and let $\text{cov}(s_0, s_1) = \sigma_{01}$.

(a) Each worker has a choice about whether to become a programmer or a lawyer, given his or her endowment of skills $(s_0, s_1)$, but no one can work as both (and no one decides not to work at all). What is a reasonable decision rule for a worker facing this problem?

**Answer:** We can probably assume that people prefer more money to less. Then, they would choose to become a lawyer (let's use $D = 1$ as an indicator for their occupational choice) when $Y_1 \geq Y_0$. This happens if and only if $\log Y_1 \geq \log Y_0$; so

$$D = 1 \iff s_1 - s_0 \geq \log(w_0/w_1)$$

(b) Suppose workers behave according to the decision rule you wrote down in (a) above. Find an expression for the fraction of workers who become lawyers.

**Hint:** The sum of two normal random variables is also normally distributed.

**Answer:** We want to find $P(D = 1)$. Given the above decision rule, this is $P(s_1 - s_0 \geq \log(w_0/w_1))$. As a shorthand, let $\Delta = s_0 - s_1$ and let $\sigma_\Delta^2 = \sigma_0^2 + \sigma_1^2 - 2\sigma_{01}$. The distribution of $\Delta$ is normal with mean zero and variance $\sigma_\Delta^2$. Then the supply of workers to the legal sector is

$$P(D = 1) = P(s_1 - s_0 \geq \log(w_0/w_1)) = P\left(\frac{\Delta}{\sigma_\Delta} \geq \frac{\log(w_0/w_1)}{\sigma_\Delta}\right) = 1 - \Phi\left(\frac{\log(w_0/w_1)}{\sigma_\Delta}\right)$$

where $\Phi$ is the standard normal CDF.

[5 + 10 = 15 points]

3. Consider the random variable $X$ with density

$$f_X(x) = \begin{cases} \frac{x}{9} & \text{if } x \in [0, 3] \\ \frac{6}{9} - \frac{x}{9} & \text{if } x \in (3, 6] \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

(a) Find the CDF of $X$, $F_X(x)$. Draw the density $f_X$ and the CDF $F_X$ on the same set of axes.

**Answer:** This is an exercise in integration. The CDF is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^2}{18} & \text{if } x \in [0, 3] \\ -1 + \frac{6}{9}x - \frac{x^2}{18} & \text{if } x \in (3, 6] \\ 1 & \text{if } x > 6 \end{cases}$$

(b) Find the inverse CDF. (The inverse CDF is the function $x = G(p)$ such that $F_X(G(p)) = p$, i.e. given a probability $p$, $G$ returns the value $x$ such that $P(X \leq x) = p$.)

**Answer:** For $0 \leq p \leq 1/2$, $G(p) = \sqrt{18p}$ solves $F_X(G(p)) = p$. For $p \in (1/2, 1]$, setting

$$p = -1 + \frac{6}{9}x - \frac{x^2}{18}$$

and solving for $x$ leads to

$$G(p) = 6 - \sqrt{36 - 18(1+p)} \text{ for } p \in (1/2, 1].$$

(We take the positive root because otherwise we'd have $x(p) < 0$, which is inconsistent with the description of the density $f_X(x)$.)

[5 + 5 = 10 points]

4. Table 1.2 of Chapter 1 of Deaton (1997), which is included in the assignment repository, contains summary statistics about the joint distribution of income and consumption for Ivorian households over the two years 1985-1986. Use the notation $y_{it}$ for the income of household $i$ in year $t$, and $c_{it}$ for the consumption of household $i$ in year $t$.

(a) What is the correlation between consumption in the two years?

**Hint:** use the fact that $V[c_{i,t+1} - c_{i,t}] = V[c_{i,t+1}] + V[c_{i,t}] - 2\text{cov}(c_{i,t+1}, c_{i,t})]$.

**Answer:** We have $V[\Delta c] = 987^2$; $V[c_{t+1}] = 1513^2$; $V[c_t] = 1236^2$. So,

$$\text{cov}(c_{t+1}, c_t) = \frac{-1}{2}\left\{987^2 - 1513^2 - 1236^2\right\}. \tag{4}$$

And, dividing by $\sqrt{V[c_t]V[c_{t+1}]} = 1513 \cdot 1236$, we get

$$\begin{aligned} \rho &= \frac{-1}{2}\left\{\frac{987^2}{1513 \cdot 1236} - \frac{1513}{1236} - \frac{1236}{1513}\right\} \\ &\approx 0.76. \end{aligned} \tag{5}$$

(b) Suppose these data come from a simple random sample of households. Further, suppose there is some measurement error in the income data, so that

$$y_{it} = y^*_{it} + \varepsilon_{it} \tag{6}$$

where $y^*_{it}$ true income, and $\varepsilon_{it}$ is "noise" that is independent of $y^*_{it}$ each period, with $E[\varepsilon_{it}] = 0$ and $V[\varepsilon_{it}] = \sigma^2$.

However, measurement error may be persistent over time (perhaps because it is driven by negligence on the part of the reporting household), so let $\rho$ be the correlation between $\varepsilon_{i,t+1}$ and $\varepsilon_{it}$.

Using the notation $\Delta y_{i,t+1} = y_{i,t+1} - y_{it}$ for the observed change in income for household $i$, and similarly for true income $y^*_{it}$, how would you calculate the variance in true income growth from panel data on $y_{it}$, if you knew $\rho$ and $\sigma^2$?

**Answer:** The variance in observed income is $V[\Delta y] = V[\Delta y^* + \Delta\varepsilon] = V[\Delta y^*] + V[\Delta\varepsilon]$, by the independence of $\varepsilon$ and $y^*$. Then, notice that $V[\Delta\varepsilon] = V[\varepsilon_{t+1}] + V[\varepsilon_t] - 2\text{cov}(\varepsilon_{t+1}, \varepsilon_t) = 2\sigma^2 - 2\sigma^2\rho = 2\sigma^2(1-\rho)$. So if we have panel data on $y$ and we know $\rho$ and $\sigma^2$, we can compute $V[\Delta y^*]$ as $V[\Delta y] - 2\sigma^2(1-\rho)$.

(c) Consider two methods of estimating the mean change in income, $E[\Delta y^*_{i,t+1}]$, over this period: (i) taking two independent cross-sectional surveys, each of size $n$, and (ii) collecting a two-period panel, also of size $n$. For method (i), say the households surveyed at time $t$ are labelled $i = 1, \ldots n$ and those surveyed at $t+1$ are labelled $i = n+1, \ldots 2n$. The first method leads to the estimator

$$\overline{\Delta} = \frac{1}{n}\sum_{i=n+1}^{2n} y_{i,t+1} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t} \tag{7}$$

while the second leads to the estimator

$$\widehat{\Delta} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \Delta y_{i,t+1} \tag{8}$$

Under what conditions - on $\sigma^2$, $\rho$, and the joint distribution of $(y^*_{i,t+1}, y^*_{i,t})$ - will the first method be more precise than the second? (Here, "precise" means having a lower variance.) Does your answer depend on the sample size? If so, why?

**Answer:** We have

$$\begin{aligned} V[\overline{\Delta}] &= n^{-1}V[y_{i,t+1}] + n^{-1}V[y_{it}] \\ &= n^{-1}(V[y^*_{t+1}] + V[y^*_t] + 2\sigma^2) \end{aligned} \tag{9}$$

where the first equality follows from the independence of the two cross-sections. The variance of the panel estimator is

$$\begin{aligned} V[\widehat{\Delta}] &= n^{-1}V[\Delta y] \\ &= n^{-1}(V[\Delta y^*] + 2\sigma^2(1-\rho)) \end{aligned} \tag{10}$$

Thus, $V[\overline{\Delta}] \leq V[\widehat{\Delta}]$ whenever

$$V[y^*_{t+1}] + V[y^*_t] + 2\sigma^2 \quad \leq \quad V[\Delta y^*] + 2\sigma^2(1-\rho) \tag{11}$$

which we can rearrange to give

$$\rho\sigma^2 \leq -\text{cov}(y^*_{t+1}, y^*_t) \quad \Longleftrightarrow \quad \rho\sigma^2 + \text{cov}(y^*_{t+1}, y^*_t) \leq 0 \tag{12}$$

i.e when the sum of the autocovariances (of the measurement error $\varepsilon_t$ and true income $y^*_t$) is less than zero, we should prefer two independent samples to a panel.[1]

To explore the intuition for this result, imagine measurement errors are independent over time, so $\rho = 0$. Then we should prefer a panel to two independent cross-sections whenever $\text{cov}(y^*_{t+1}, y^*_t) > 0$, i.e true incomes are persistent (which is typically the case). This is because the cross-sectional variation in $y^*$ represents mainly permanent differences across households (and measurement error). Thus, the changes in income are less variable than the levels.

On the other hand, if $\rho < 0$ so the measurement error is mean-reverting, repeated observations on the same households are less useful for measuring changes. In this case, the observed changes in $y$ will tend to be dominated by fluctuations in measurement error rather than real changes in $y^*$.

[5 + 5 + 10 = 20 points]

# Data Manipulation in R

## Filtering, Sorting, and Generating New Variables

5. Do exercise 1 from section 5.2.4 of Grolemund and Wickham (2017).

[5 points]

6. Do exercises 3-4 of section 5.3.1 of Grolemund and Wickham (2017).

[1] You didn't need to prove this, but you should confirm that both estimators are *unbiased*, meaning that their expected values are $E[\Delta y^*]$.

[$2 \times 5 = 10$ points]

7. Do exercise 2 of section 5.5.2 of Grolemund and Wickham (2017)

[5 points]

### Grouped Summaries and Filters

8. Do exercise 6 of section 5.7.1 of Grolemund and Wickham (2017)

[5 points]

9. Do exercise 4 of section 5.6.7 of Grolemund and Wickham (2017)

[5 points]

### Reshaping Data

10. Do exercise 2-4 of section 12.3.3 of Grolemund and Wickham (2017)

[$3 \times 5 = 15$ points]

## Counterfactuals

Included in this repository are six short texts. Read each one carefully, and, identify the claims being made. Classify them as either causal claims, or non-causal claims. If they are not causal, what are they - value judgements? A statement of fact? Something else?

For those claims you think are causal, comment on whether the authors present any evidence for their claim. If so, what might be some concerns or alternative interpretations of that evidence? If they do not present any evidence for their claims, what sort of information would you need to decide whether it was true or false?

**Please note:** You are welcome to have any views you like about these topics, but I will deduct points for extraneous opinion. The point of this exercise is to read the texts carefully and critically, and to be able to recognize causal statements (including implicit ones). It is **not** to discuss the particular issues mentioned in the texts.

**Answer(s):**

*Schussler EWC tweet*

- This is a causal claim (although posed as a question).
- The evidence presented is the aggregate time series of completed building construction. One obvious problem with the picture (taken in isolation) is that the series is very volatile anyway (e.g. look at the drop in mid-2009).
- At these fairly high frequencies, it might have been better to look at building starts, i.e. the initiation of new construction. Some of the drop in completed building happening in the present might be because new construction slowed down some months or years ago.
- The ideal experiment might have been to randomly assign expropriation risk to different but comparable populations.
- Overall it's hard to tell whether Schussler's claim is true or not, at least given this evidence.

*Ramakoa tweet*

- The first statement about SA business' "instincts" is a value judgement.
- The second statement might be causal, although it's unclear what the implied counterfactual is. If "capital systematically distorts labour pricing", we have to ask relative to what?

  - An economy where there is no capital whatsoever?
  - An economy with less capital per worker?
  - Something else?
- The author seems to be suggesting that firms have a lot of wage-setting power, although what the wage distribution could (and in the author's view "should") be isn't clear.
- We could also ask whether "distortions" of the wage distribution always increase inequality. It could be that the "undistorted" wage distribution would be even more dispersed than the one we see in reality.

*Landlessness tweet*

- The first, second, and third statements are factual claims.
- For the third statement, the "dop system" of paying farmworkers in the form of alcohol is certainly a well-known phenomenon, although I have never seen documentation of exactly how common it is.
- There is an implicit causal claim that landlessness is the cause of high rates of FAS in the rural Western Cape, but no evidence is presented for it.
  - We might ask whether it is land itself, or wealth more generally, that is the cause of high FAS rates. (I would guess that there are many poor communities in, e.g. Eastern Europe where alcoholism and FAS are common, but people own the land they work.)
  - It's also possible to interpret "landlessness" more loosely, as a metonym for colonialism or even poverty.

*Minimum wage tweets*

- Most of these claims are factual - either about aspects of the wage distribution or about the legal details of the proposed act.
- There are also some value judgements, e.g. R20/hour is "not enough".
- Most counterfactuals here can be interpreted "mechanically", i.e. they assume no behavioral response by workers or firms.
- There is a causal claim implied by the statement "It is not feasible for wages to jump massively in one go." (Why not? What would happen?)
- The final statement - that the aggregate effect of raising the national minimum wage would be to raise both the level of output and the rate of growth - is causal. The author gives no evidence for it, though.

*Makgetla article*

- There are some factual claims made about the distribution of income across different types of statistical units (individual persons, households, and firms).
- There is a causal claim made that, if the size distribution of firms is right-skewed, then the managers and shareholders of the largest firms can make decisions that "largely shape the economy".
  - This is vague because the author does not say in which respects these managers and shareholders can "shape the economy". For example, can they affect education levels or consumer debt levels?
  - It is also unclear what is being assumed about the reactions of other firms or consumers and workers to these possible decisions. For example, could the managers and the shareholders of the largest firms unilaterally set wages for all workers?
- The main problem with the causal claims in this article are that it is vague about what "power" is and how it can be exercised in practice, so it is hard to evaluate its claims.
  - For example, the author mentions that some asset managers have very large portfolios. But these are liabilities to those asset managers!
  - The author seems to be implying that asset managers are completely free to allocate them, which seems unlikely unless investors are completely passive.
  - There may be other interpretations consistent with the author's words and with her intended meaning.

*Vox luxury gyms article*

- The first part of the article describes differences in exercise patterns and membership of fitness clubs by income. These are factual claims.
- There is a causal innuendo (boutique fitness chains "play into" exercise inequality).

  - The suggestion seems to be that these firms are exercising discretion by serving an affluent set of consumers, but they could choose to do otherwise.
  - You often read newspaper articles along the same lines about how rich consumers buy healthier groceries than poor consumers, and high-end food shops don't locate in poor neighborhoods.
  - There is an obvious reverse-causality story here: these firms locate in rich neighborhoods because that's where their customers are, and they don't locate in poor neighborhoods because it would not be profitable to do so. The authors do not entertain this possibility.
- Then, the article claims that poor people on average have less access to public parks and other exercise facilities. This is a factual claim.
- Whether this difference in access to public parks is a cause of different exercise patterns by income does not follow from this fact, although the article does seem to suggest so.
  - Even if richer people have greater access (say by distance) to public parks, the starting observation of this article is that the rich are exercising in expensive *private* gyms.
  - We would want to know how much exercise rich people do in public parks compared to private gyms in order to see if their greater access to parks is the cause of their higher exercise levels.

[$6 \times 10 = 60$ points]

# References

Grolemund, Garrett, and Hadley Wickham. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly. https://r4ds.had.co.nz.