# Week 2 Handout

Gov 50 Data Science for the Social Sciences

John Koo

September 9, 2025

## Math Recap

### Probabilities and expectations

| | Discrete Variables | Continuous Variables |
|---|---|---|
| Distribution | | |
| Probability | $\Pr(X = x) = \frac{n_x}{n}$ | $f(a < x < b) = F(b) - F(a)$ |
| Expectation | $\mathbb{E}[X] = \sum_x x \Pr(X = x)$ | $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$ |

*Q: What is the expectation of a six-sided die?*

Let $X$ be the number on the die,

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot \Pr(X = 1) + 2 \cdot \Pr(X = 2) + \cdots + 6 \cdot \Pr(X = 6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} \\ &= \frac{21}{6} \\ &= 3.5\end{aligned}$$

## Conditional Probabilities and Conditional Expectation

Bayes Rule:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

*Q: What is the expectation of a six-sided die, given that the number is even?*

$$\begin{aligned}\mathbb{E}[X \mid \text{even}] &= 2 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} \\ &= \frac{12}{3} \\ &= 4\end{aligned}$$

## Potential Outcomes Framework and Causal Inference

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

$$\widehat{\text{ATE}} = \text{SDO} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

(Note: Hat means "estimator of"; vanilla ATE is not identifiable IRL because of the fundamental problem of causal inference - you can't observe $Y_i(1)$ and $Y_i(0)$ at the same time)

$$\text{ATT} = \mathbb{E}[Y_i(1) \mid D_i = 1] - \underbrace{\mathbb{E}[Y_i(0) \mid D_i = 1]}_{\text{not observable}}$$

$$\text{ATU} = \underbrace{\mathbb{E}[Y_i(1) \mid D_i = 0]}_{\text{not observable}} - \mathbb{E}[Y_i(0) \mid D_i = 0]$$

## Class practice

### Introduction

Are democracies better for economic development than nondemocracies (autocracies)? Could having a democratic regime be related to economic growth? These questions have fascinated social scientists, policy-makers and pundits for many decades now. In a recent publication, Acemoglu, Naidu, Restrepo and Robinson (2019) describe the debate in the following terms:

"With the spectacular economic growth under nondemocracy in China, the eclipse of the Arab Spring, and the recent rise of populist politics in Europe and the United States, the view that democratic institutions are at best irrelevant and at worst a hindrance for economic growth has become increasingly popular in both academia and policy discourse. For example, the prominent New York Times columnist Tom Friedman (2009) argues that 'one-party nondemocracy certainly has its drawbacks. But when it is led by a reasonably enlightened group of people, as China is today, it can also have great advantages. That one party can just impose the politically difficult but critically important policies needed to move a society forward in the 21st century'. Robert Barro (1997, 1) states this view even more boldly: 'More political rights do not have an effect on growth.' "

In this exercise we'll take a stab at this debate by answering the following related question: are democracies richer than autocracies?

We will be working with the dataset from the paper by Acemoglu, Naidu, Restrepo and Robinson (2019). The dataset includes the following variables:

| Name | Description |
|---|---|
| `country_name` | Country name |
| `wbcode` | World Bank country code |
| `year` | Year |
| `gdppc` | GDP per capita (constant 2000 US$) |
| `region` | Geographical region |
| `dem` | Democracy measure (1 = Democracy; 0 = Autocracy) |

### Question 1: Loading the dataset

Before we can get started working with data, we first need to load the data into R. Datasets can come in many file types, but the most common is a CSV, which stands for "comma-separated values". Use the `read.csv()` function from the R package `readr` to read your data into R and call it `anrr`. You'll find the dataset under the folder `data`. This is the original data used in their study.

### Answer 1

```
library(readr)

anrr <- read_csv("data/section1_df.csv",
                 show_col_types = FALSE)
```

### Question 2: Inspecting the dataset I

Use the `head()` function to view the first several rows of the data. What can you notice about the variable gdppc?

### Answer 2

```
head(anrr)
```

```
## # A tibble: 6 x 7
##    ...1 country_name wbcode  year gdppc region   dem
##   <dbl> <chr>        <chr>  <dbl> <dbl> <chr>  <dbl>
## 1     1 Afghanistan  AFG     1960    NA MNA        0
## 2     2 Afghanistan  AFG     1961    NA MNA        0
## 3     3 Afghanistan  AFG     1962    NA MNA        0
## 4     4 Afghanistan  AFG     1963    NA MNA        0
## 5     5 Afghanistan  AFG     1964    NA MNA        0
## 6     6 Afghanistan  AFG     1965    NA MNA        0
```

There are 9384 country-years with missing values in the GDP per capita variable.

### Question 3: Inspecting the dataset II

Use the function `glimpse()` to look at a summary of the dataset. What can you notice about the variable dem?

### Answer 3

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(anrr)
```

```
## Rows: 9,384
## Columns: 7
## $ ...1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ country_name <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan~
## $ wbcode       <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "~
## $ year         <dbl> 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 196~
## $ gdppc        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ region       <chr> "MNA", "MNA", "MNA", "MNA", "MNA", "MNA", "MNA", "MNA", "~
```

```
## $ dem          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

The variable `dem` is numeric (a double).

### Question 4: Measuring democracy

There are many potential ways to code if a country is democratic or not. Researchers came up with criteria to classify political regimes as a binary variable (if you are interested how, check Boix, Miller and Rosato 2012). In these measurements, democracies are often coded as a 1, and nondemocracies (autocracies) as a 0.

Use the function `table` to see how many observations are democracies and how many autocracies in the data. (Hint: to tabulate the values of a variable, pass as arguments to `table` something of the form `dataframe$variable`).

### Answer 4

```r
table(anrr$dem)
```

```
##
##    0    1
## 4956 3777
```

### Question 5

Now add as an argument to the function `table` the option `useNA = "always"`. How many missing values does the variable `dem` has?

### Answer 5

```r
table(anrr$dem, useNA = "always")
```

```
##
##    0    1 <NA>
## 4956 3777  651
```

### Question 6

When we create data visualizations, we sometimes want to make numeric variables like `dem` a `factor`. In this way, we can acknowledge that, although imported as numeric, the variable represents two distinct categories: democracy and autocracy. Run the following code:

```r
library(dplyr)

anrr <- anrr |>
  mutate(dem_label = factor(dem,
                            levels = c(1, 0),
                            labels = c("Democracy", "Autocracy")))
```

Since we are only comparing democracies and autocracies, we can also leave aside the `NA`s for visualization sake.

```
anrr <- anrr |>
  filter(!is.na(dem))
```

Check the class of the new variable and corroborate that it shares the same number of democracies and autocracies, but that we have no NAs.

### Answer 6

```
anrr <- anrr |>
  mutate(dem_label = factor(dem,
                            levels = c(1, 0),
                            labels = c("Democracy", "Autocracy")))

anrr <- anrr |>
  filter(!is.na(dem))
```

```
class(anrr$dem_label)
```

```
## [1] "factor"
```

```
table(anrr$dem_label, useNA = "always")
```

```
##
## Democracy Autocracy      <NA>
##      3777      4956         0
```

### Question 7: Visualizing the income distribution

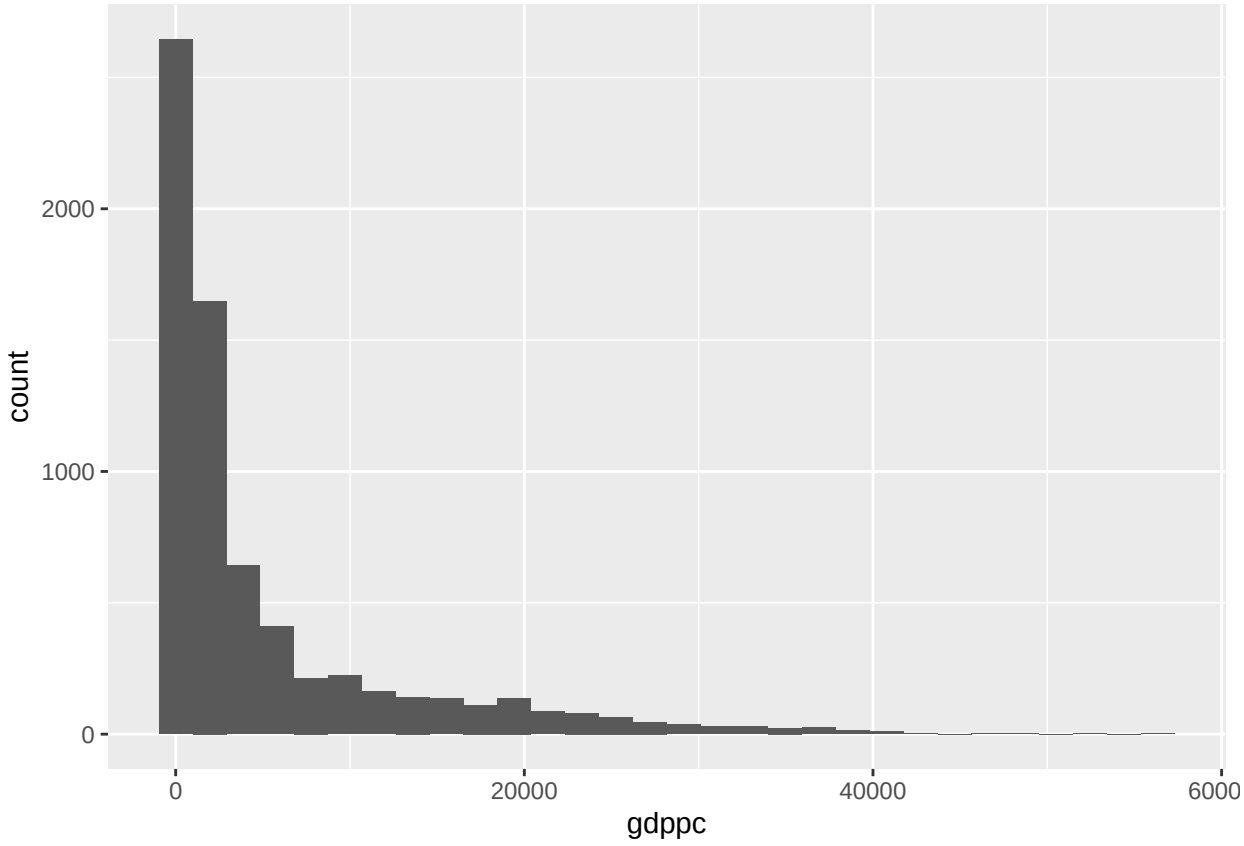
Now we can start comparing how rich (or poor) are democracies relative to autocracies.

Using ggplot, plot an histogram of the variable gdppc. What can you say about the distribution of the variable?

### Answer 7

```
library(ggplot2)

ggplot(anrr, aes(x=gdppc)) +
  geom_histogram()
```

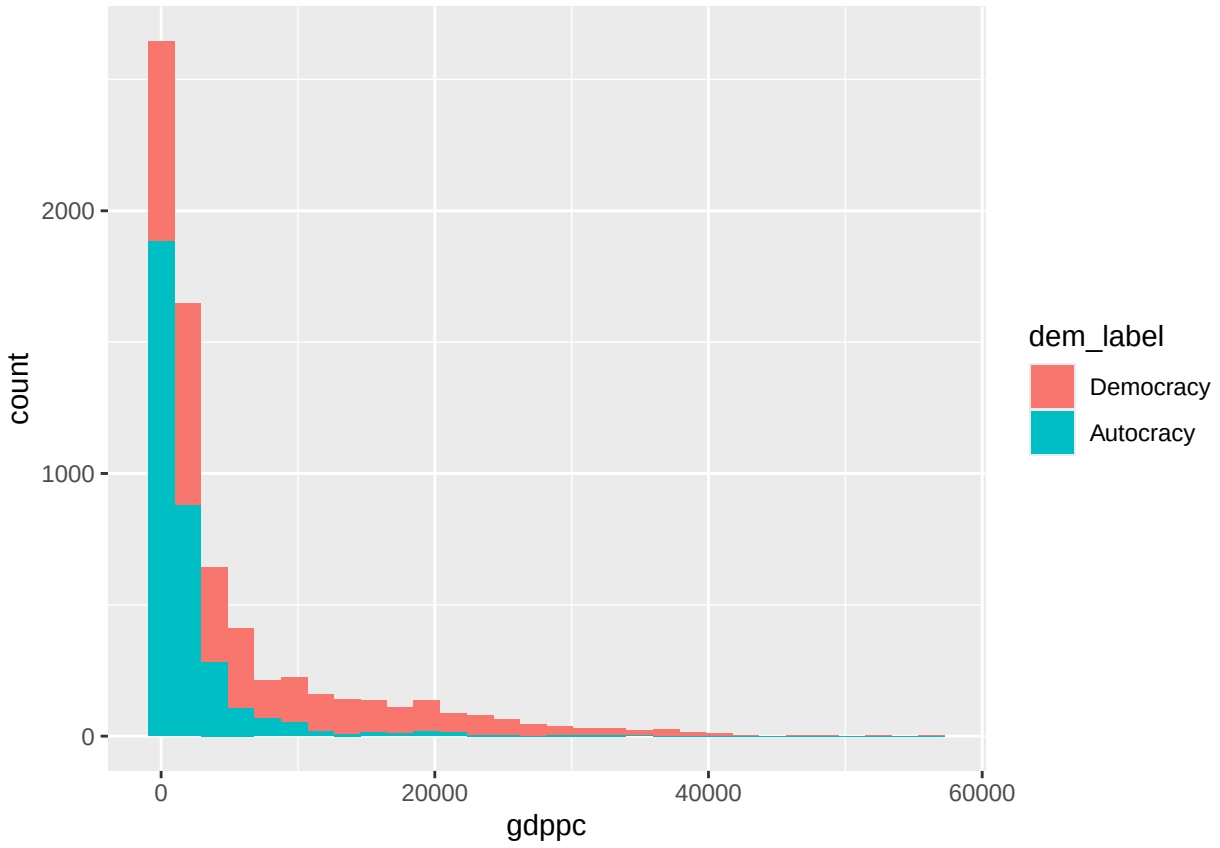



### Question 8: Comparing income by regime

The plot above is showing the distribution of income for both democracies and autocracies together. Pass the argument `fill = dem_label` to the `aes()` to see how the distributions differ by political regime. From this plot, can you tell if democracies are richer or poorer than autocracies?

### Answer 8

```
ggplot(anrr, aes(x=gdppc, fill = dem_label)) +
  geom_histogram()
```

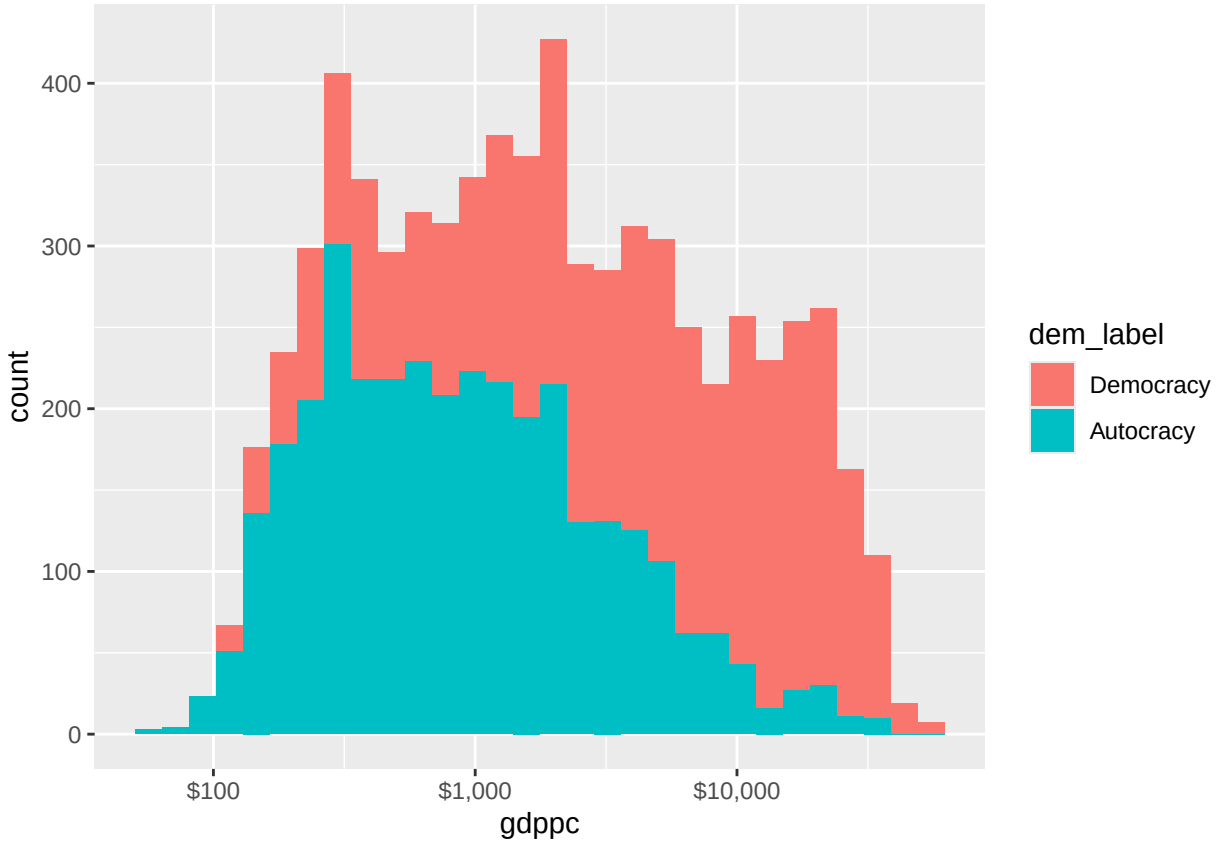



### Question 9: Log scale

When the distribution of our variable is highly skewed, we often transform the variable by a logarithmic scale. Make the same plot as in 2 above, but adding `scale_x_log10(labels = scales::dollar)` as an argument to the `ggplot`. Does this transformation makes clearer the income differences between democracies and autocracies?

### Answer 9

```
ggplot(anrr, aes(x=gdppc, fill = dem_label)) +
  geom_histogram() +
  scale_x_log10(labels = scales::dollar)
```

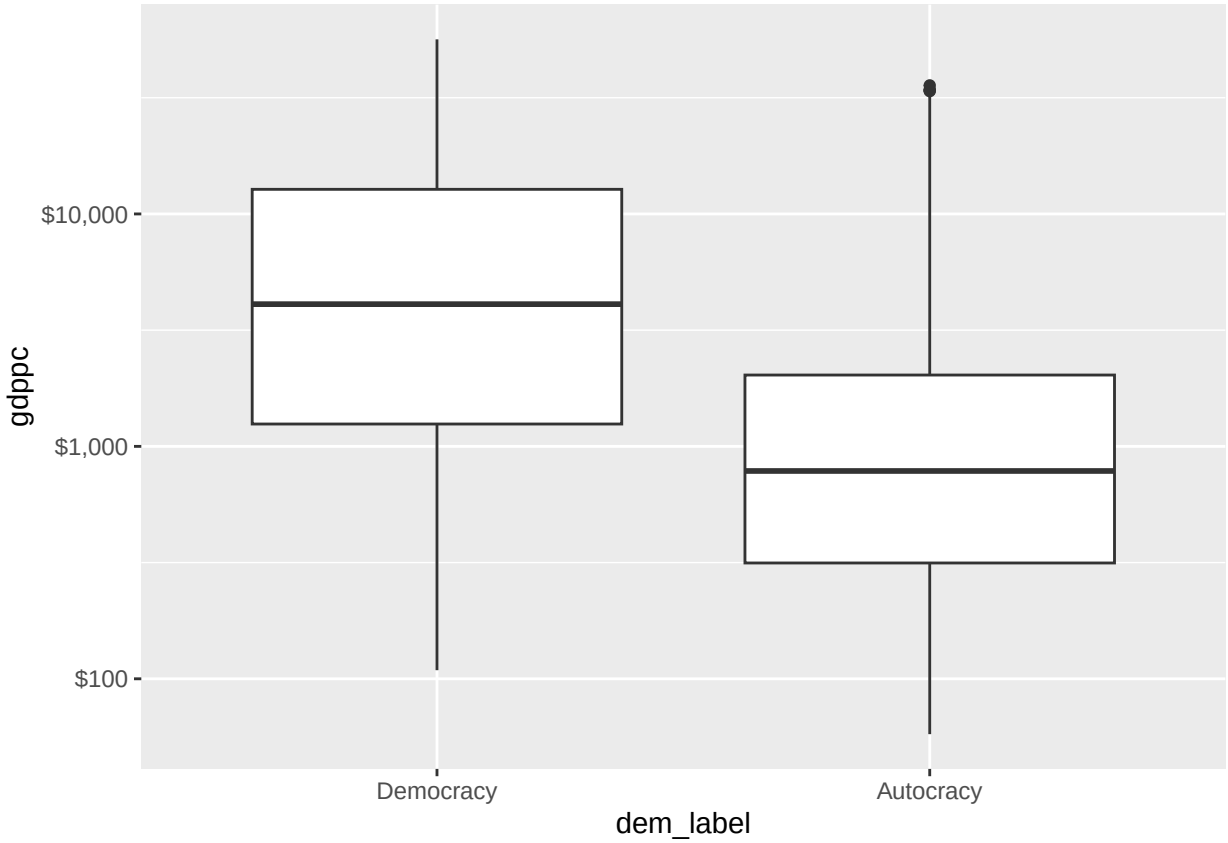



### Question 10: Comparing with boxplot

An alternative way to get to the question is to compare the median income between political regimes. We can do this using `geom_boxplot`. What does this plot tells you about the distribution of income in democracies when compared to autocracies?

### Answer 10

```
ggplot(anrr, aes(x=dem_label, y=gdppc)) +
  geom_boxplot() +
  scale_y_log10(labels = scales::dollar)
```

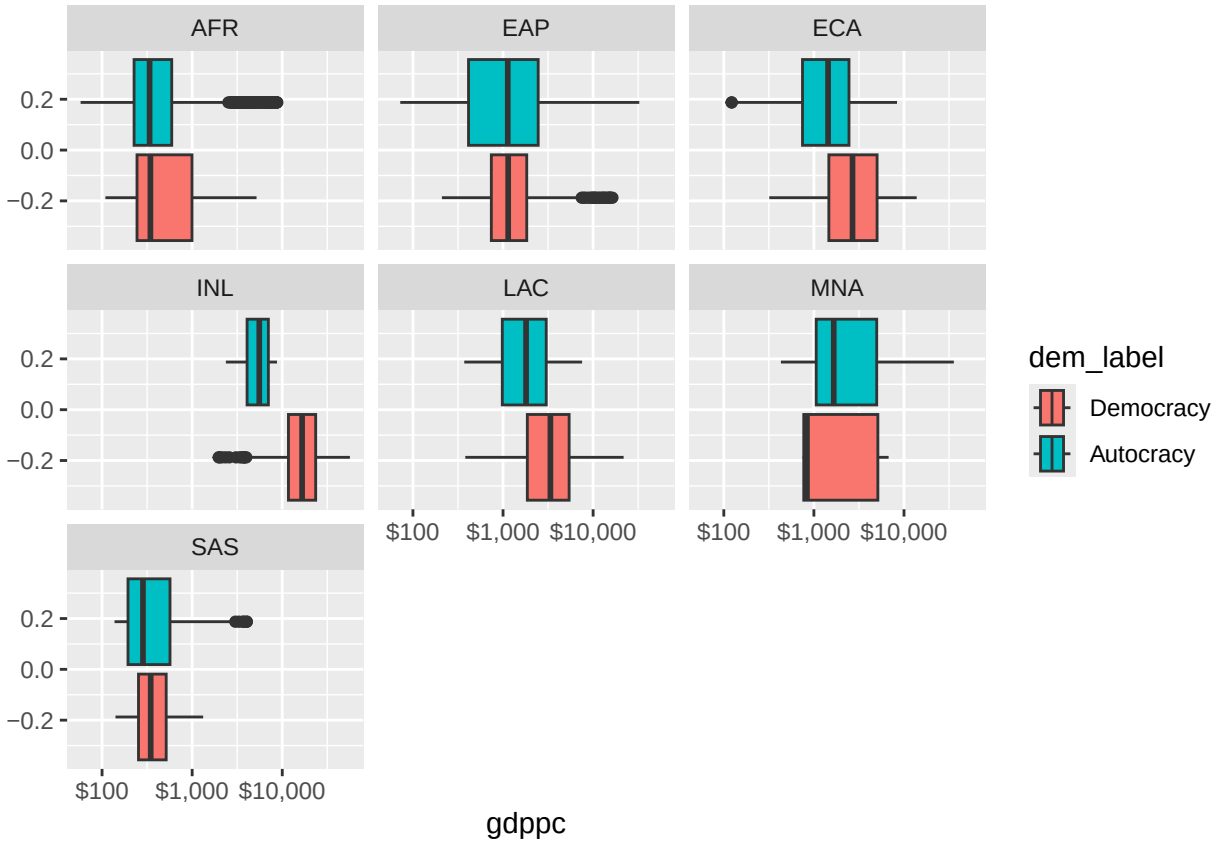



### Question 11: Comparing by country groups

Finally, do these patterns vary by world region? Add `facet_wrap(~region)` to your `ggplot` to see the division by geographic/economic region. Region acronyms are AFR: Africa , EAP: East Asia and the Pacific, ECA: Europe and Central Asia, INL: OECD and high income countries, LAC: Latin America and Caribbean, MNA: Middle East and North Africa, SAS: South Asia. Can you find a region of the world where the median income of autocracies is higher than that of democracies?

### Answer 11

```
ggplot(anrr, aes(x=gdppc, fill = dem_label)) +
  geom_boxplot() +
  scale_x_log10(labels = scales::dollar) +
  facet_wrap(~region)
```


AFR
EAP
ECA
INL
LAC
MNA
SAS
0.2
0.0
−0.2
$100
$1,000
$10,000
gdppc
dem_label
Democracy
Autocracy