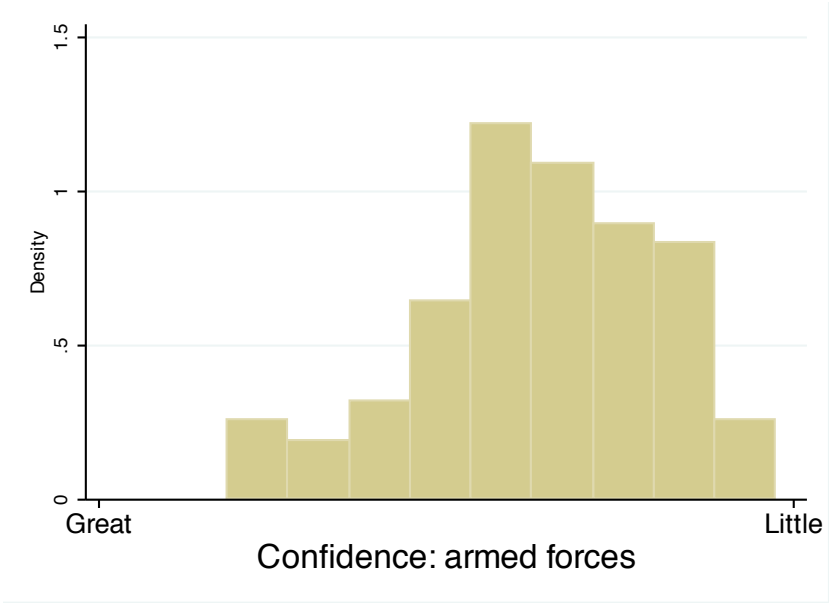
Distributions



Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 4

Launch your research project!

Fill in the details at http://goo.gl/brYmB

- Form a pair in your own group.
- Write up a clear title.

If you want to make any background reading on your topic, now is the right time to do so. Download and start editing the paper template at http://goo.gl/7u8oa if you need a writing guide.

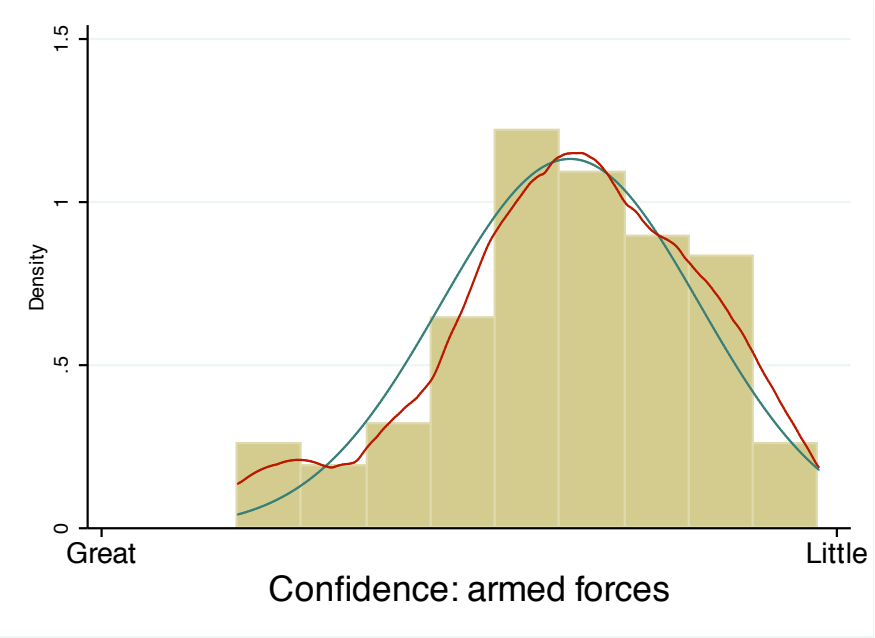
Outline

Univariate statistics look at the distribution of one variable and are part of descriptive statistics.

Central tendency

Variability

Normal distribution



Mean

The arithmetic mean $\bar{X}$ (or "x bar") of variable X with N observations is given by the following formula:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_N}{N} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Mean values can be calculated in several ways, but arithmetic means (averages) are by far the most common.
- Mean values are not robust to extreme values: their values are very sensitive to outlier observations.
- If $X_1, X_2, ..., X_N$ are all weighted by a coefficient w_i, the arithmetic weighted mean is given by $\sum_{i=1}^{N} w_i \cdot X_i$, with $\sum_{i=1}^{N} w_i = 1$.

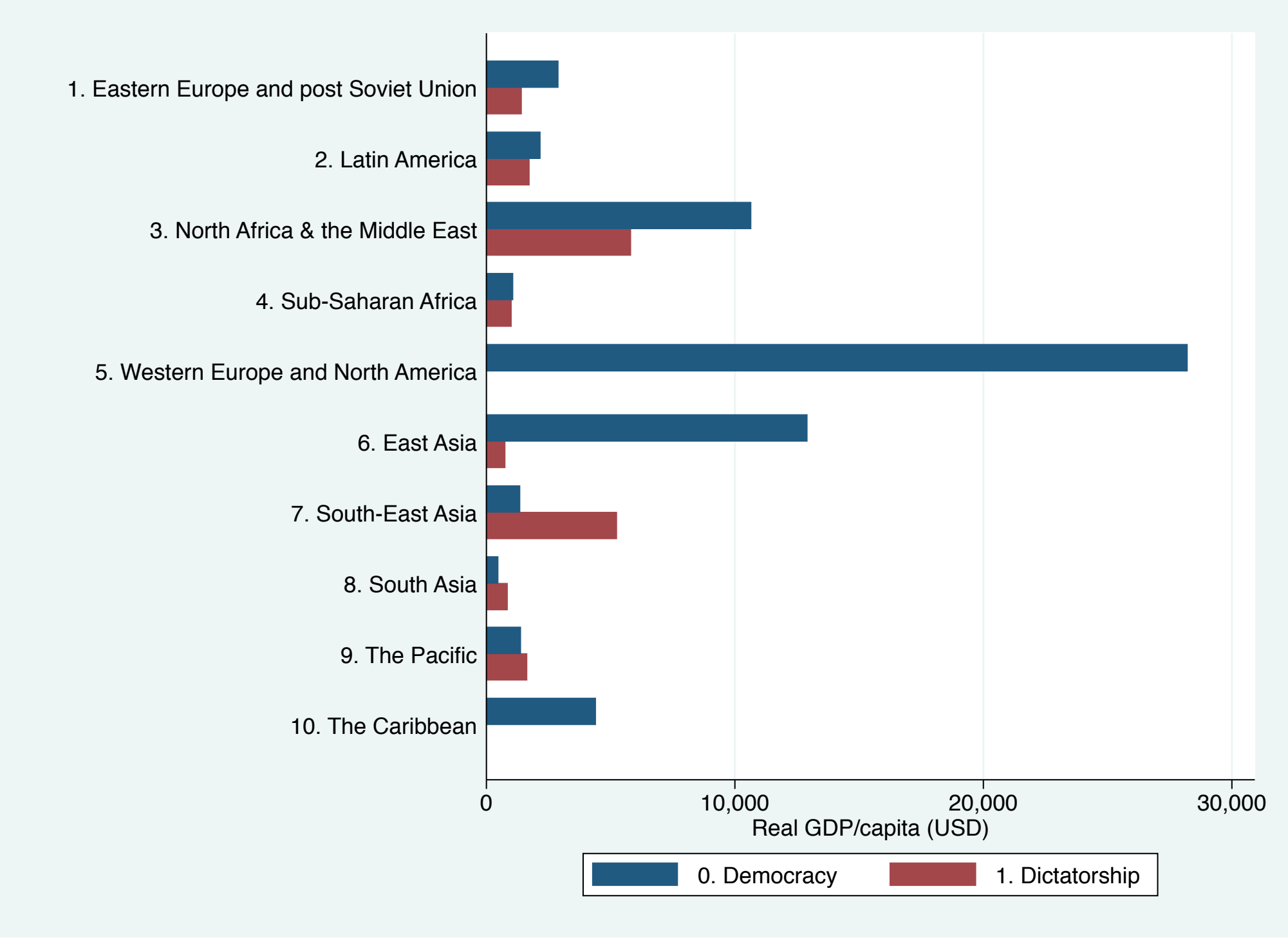
Example: GDP/capita

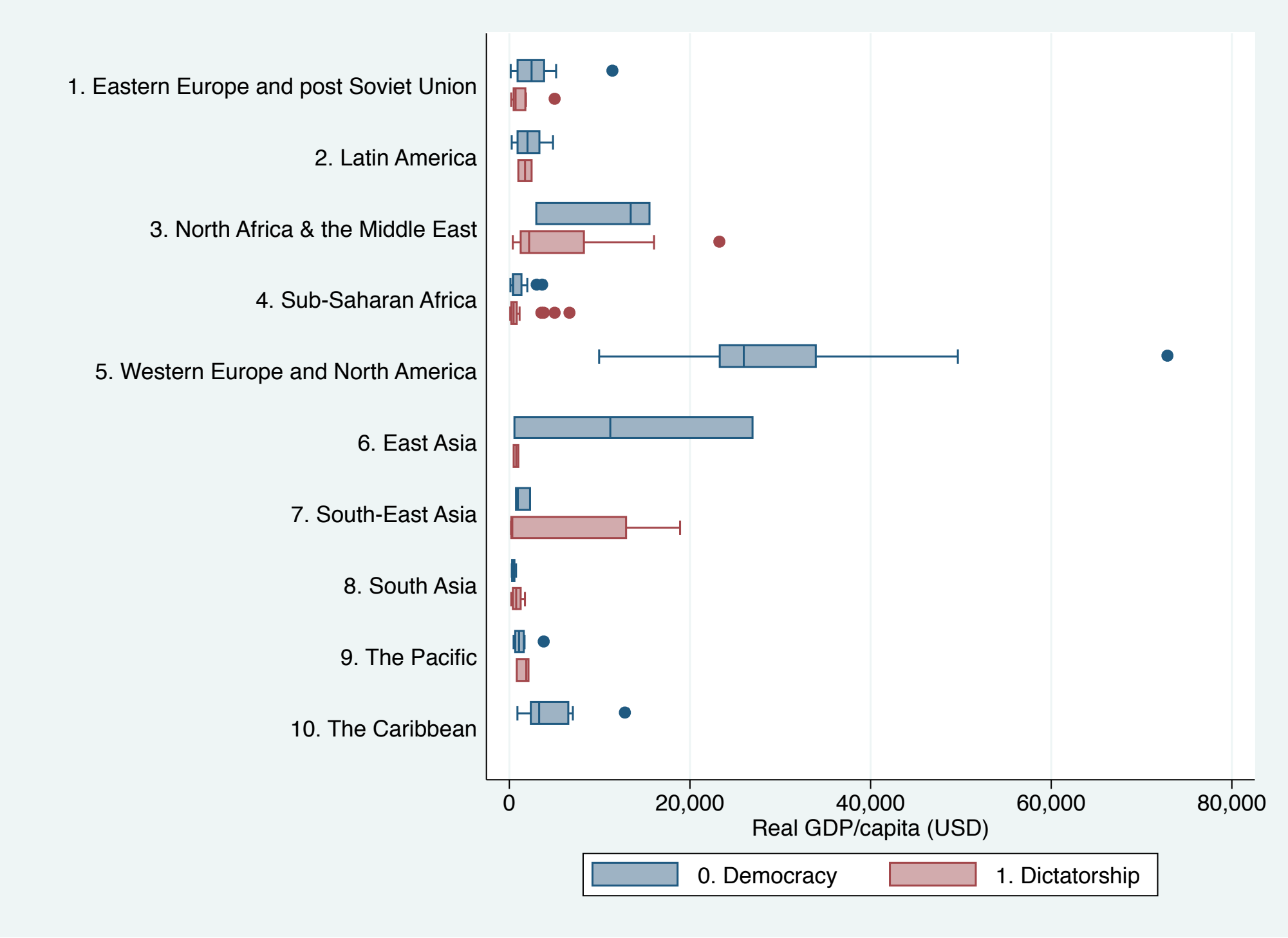
Using country-level real GDP/capita, measured by the United Nations Statistics Divisions – National Accounts in 2009:

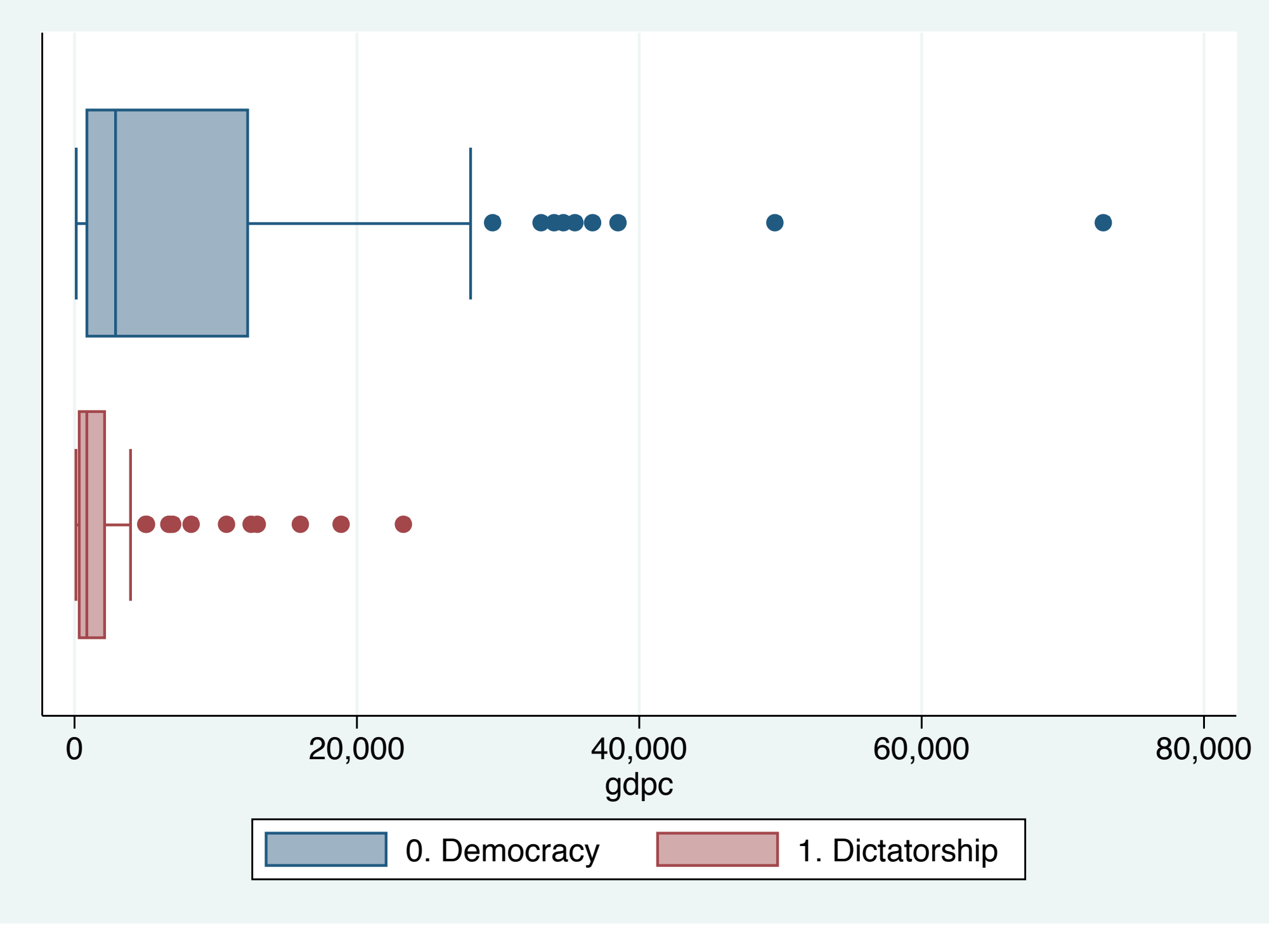
$$\text{Real GDP/capita} = \frac{\text{Real GDP}}{\text{population}} = \frac{\text{GDP}}{\text{population}} \cdot \text{price index}$$

- The mean can summarise the distribution of real GDP/capita in the full sample ($N = 192$) and/or in each geographical region.
- Average real GDP/capita will be sensitive to exceptionally high or low values, as with Somalia, Liechtenstein or Switzerland.
- Population and price act like country-level weights: all values of GDP are weighted by $\frac{\text{price index}}{\text{population}}$ to make them comparable.

On relative pricing, see Feinstein and Thomas, Appendix B.


1. Eastern Europe and post Soviet Union
2. Latin America
3. North Africa & the Middle East
4. Sub-Saharan Africa
5. Western Europe and North America
6. East Asia
7. South-East Asia
8. South Asia
9. The Pacific
10. The Caribbean
0
10,000
20,000
30,000
Real GDP/capita (USD)
0. Democracy
1. Dictatorship


1. Eastern Europe and post Soviet Union
2. Latin America
3. North Africa & the Middle East
4. Sub-Saharan Africa
5. Western Europe and North America
6. East Asia
7. South-East Asia
8. South Asia
9. The Pacific
10. The Caribbean
0
20,000
40,000
60,000
80,000
Real GDP/capita (USD)
0. Democracy
1. Dictatorship


0
20,000
40,000
60,000
80,000
gdpc
0. Democracy
1. Dictatorship

Median

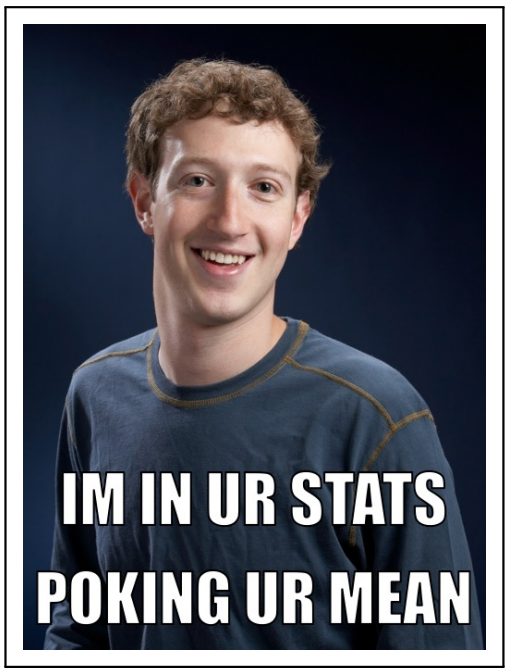
The median value is the "middle" of the distribution:

- 50% of the values fall below the median.
- 50% of the values fall above the median.

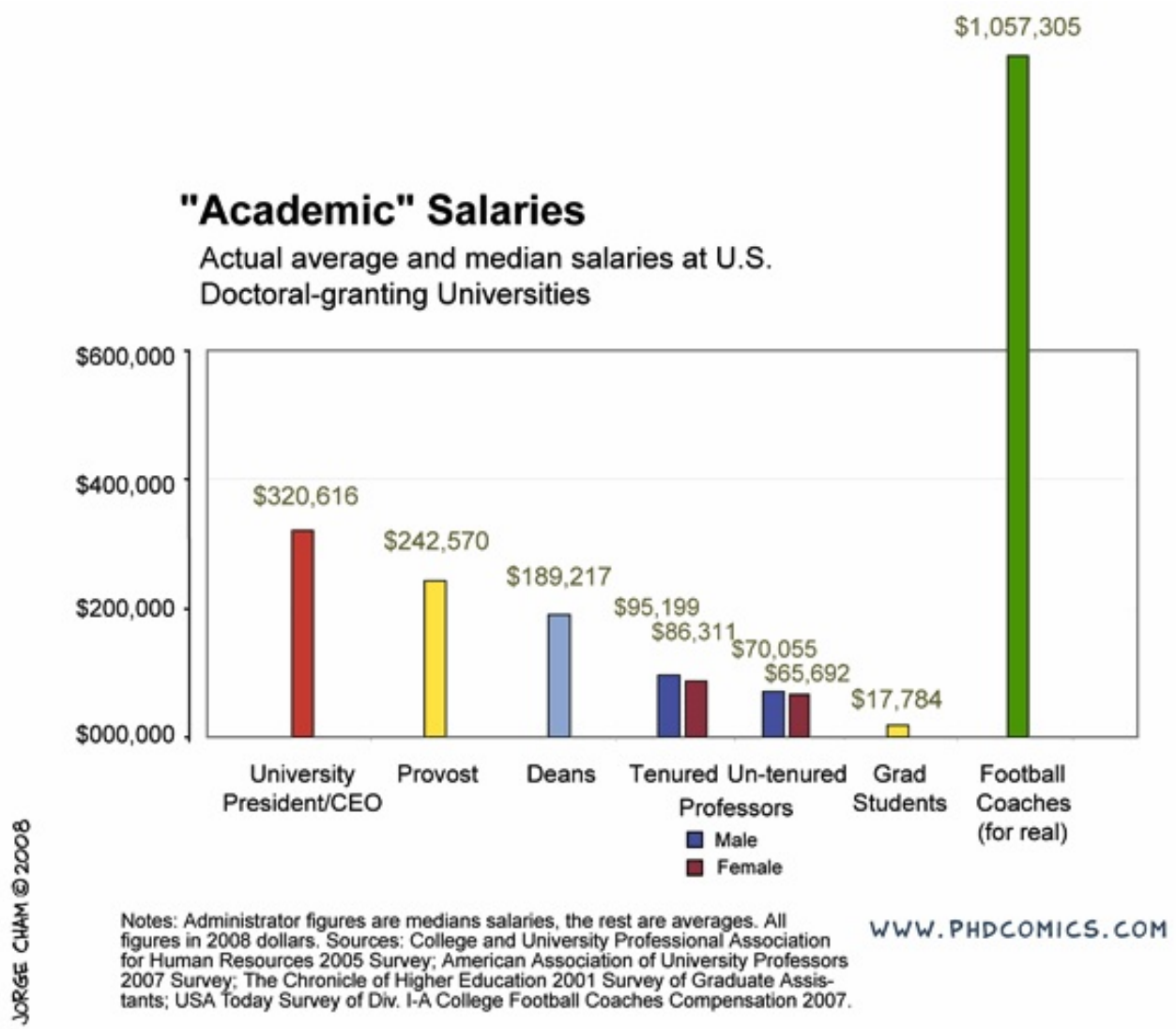
Unlike the mean, it is robust to extreme values:

- Population: young Westerners, sample: class.
- Estimate values for mean and median income.
- Enter Facebook CEO. Recalculate estimates.
- Which value has just become very misleading?

In your work, adjust your reported measures of central tendency and select among graph types after looking at both values for the variable.



Problem: median vs. mean



True!

Shockingly true!

Yet:

"What's the median salary of these football coaches?"

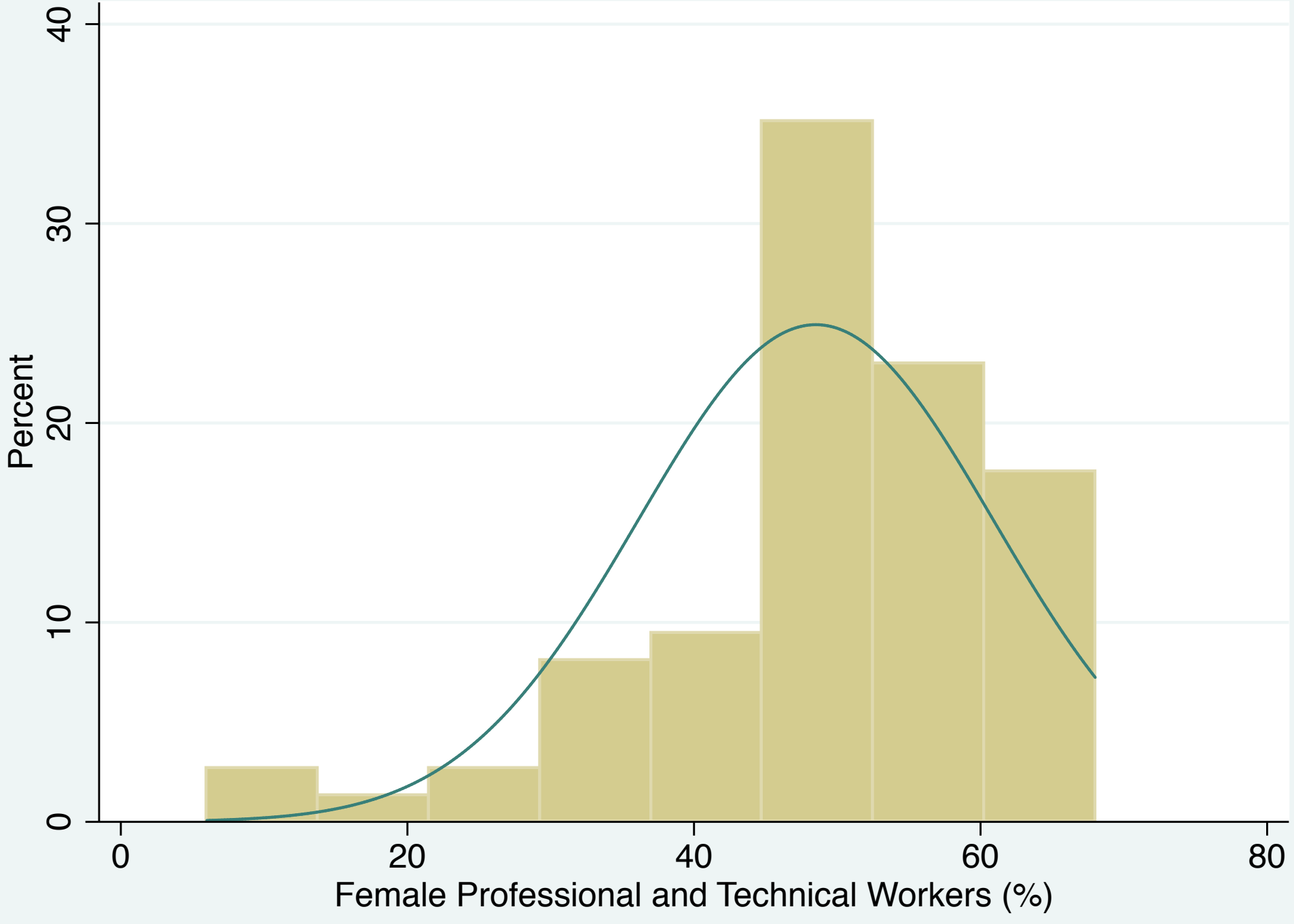
Result:

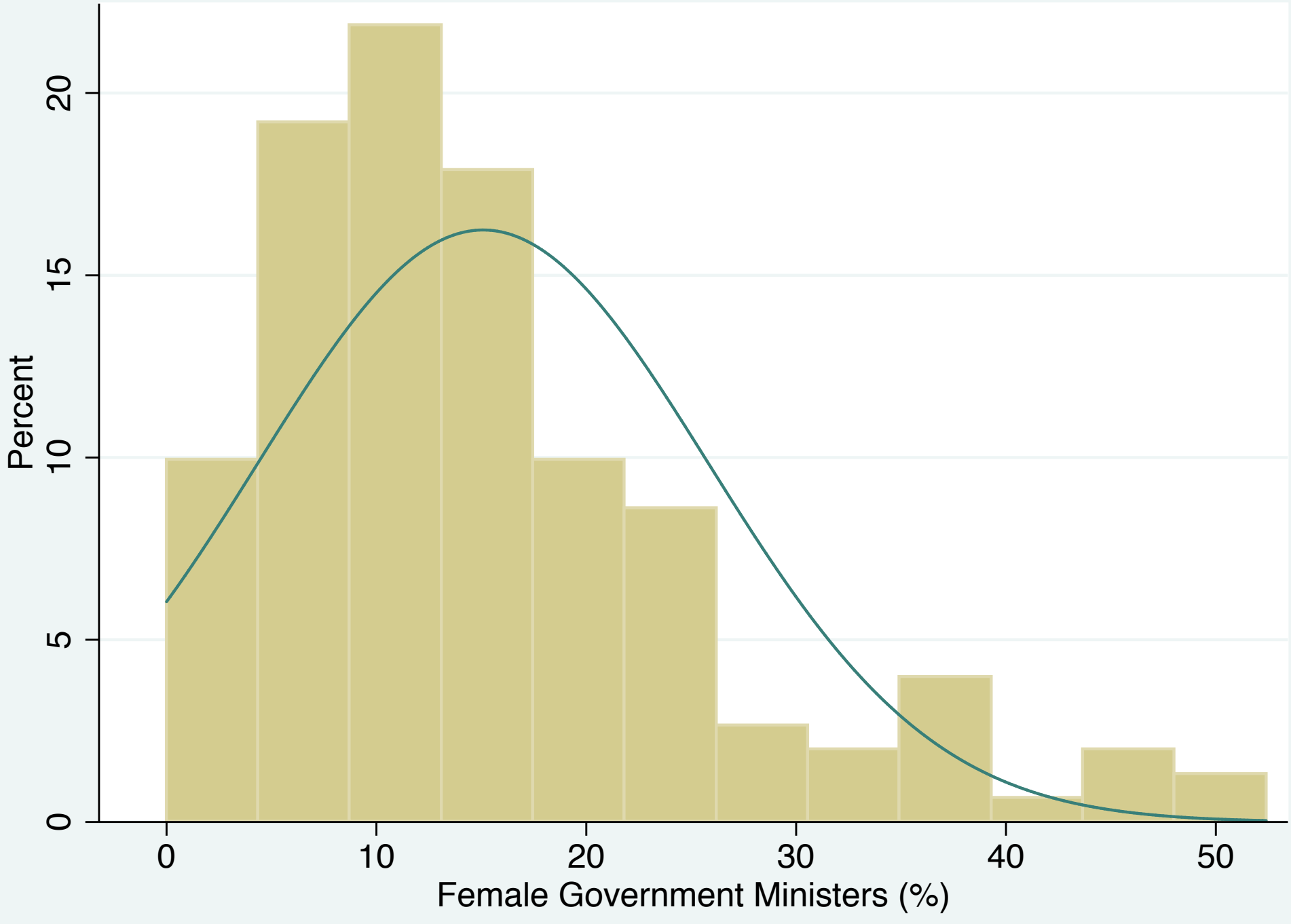
Junk charts.

Solution: skewness

Skewness measures the symmetry of the distribution by looking at the relative positions of the median and the mean:

- If median $>$ mean, the distribution comes with a "longer left tail" and shows "right-side skewness".
- If median $<$ mean, the distribution comes with a "longer right tail" and shows "left-side skewness".

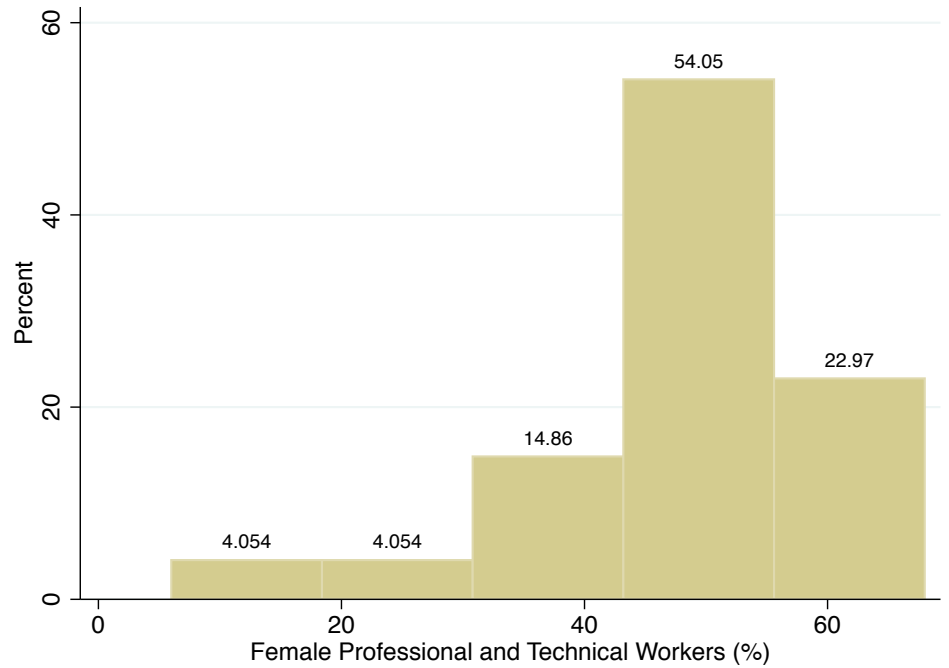

Percent
0
10
20
30
40
0
20
40
60
80
Female Professional and Technical Workers (%)


Percent
0
5
10
15
20
0
10
20
30
40
50
Female Government Ministers (%)

Mode

The mode is the 'peak' of the distribution at its most frequent value.

Histograms show the density of a distribution, to observe the mode and skewness.



Stata: `su`, `tab`, `hist` etc.

- **Formal** exploration:
  - `su` (`summarize`) provides a five-number summary for the distribution of continuous variables.
  - `tabstat` and `su` (`summarize`) with the `d` (`detail`) option provide percentiles and call the median `p50` or `50%`, for "50th percentile."
  - `su` (`summarize`) with the `d` (`detail`) option also provides skewness, between -1 and $+1$, with 0 indicating symmetry.
  - `tab` or `fre` provide frequency tables with (cumulative and valid) percentages for categorical variables, to which only the mode applies.
- **Visual** exploration:
  - `hist` (`histogram`) shows the density of the distribution.
  - `percent` and `addl` optionally add percentage scale and labels.
  - `gr dot` is the most useful plot to visualize across categorical variables.

Stata implementation

- Explore the data with `lookfor` and `d`.
- Use `su` to summarise all variables at once.
- Use `su` with `d` on the dependent variable.
- Export a table with a `tabstat` sequence.

```
. su gid_fgm gid_fptw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gid_fgm	151	15.08212	10.72527	0	52.4
gid_fptw	74	48.5	12.40222	6	68

```
. su gid_fptw, d
```

Female Professional and Technical Workers (%)

	Percentiles	Smallest		
1%	6	6		
5%	25	12		
10%	32	15	Obs	74
25%	45	25	Sum of Wgt.	74
50%	51		Mean	48.5
		Largest	Std. Dev.	12.40222
75%	55	66		
90%	61	66	Variance	153.8151
95%	66	67	Skewness	-1.192047
99%	68	68	Kurtosis	4.767732

Some descriptive measures will appear in the "summary statistics" table of your research, produced with a `tabstat` command sequence documented in the Stata Guide, Section 9.

Variability

- **Range** is the 'spread' of the variable X between its maximum and minimum values:

$$X_{max} - X_{min}$$

- **Variance** is the sum of deviations from the mean, $X_i - \bar{X}$, for each value taken by the variable X:

$$\sigma^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

- **Standard deviation** is the compound square root of variance divided by sample size N:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

Quantiles

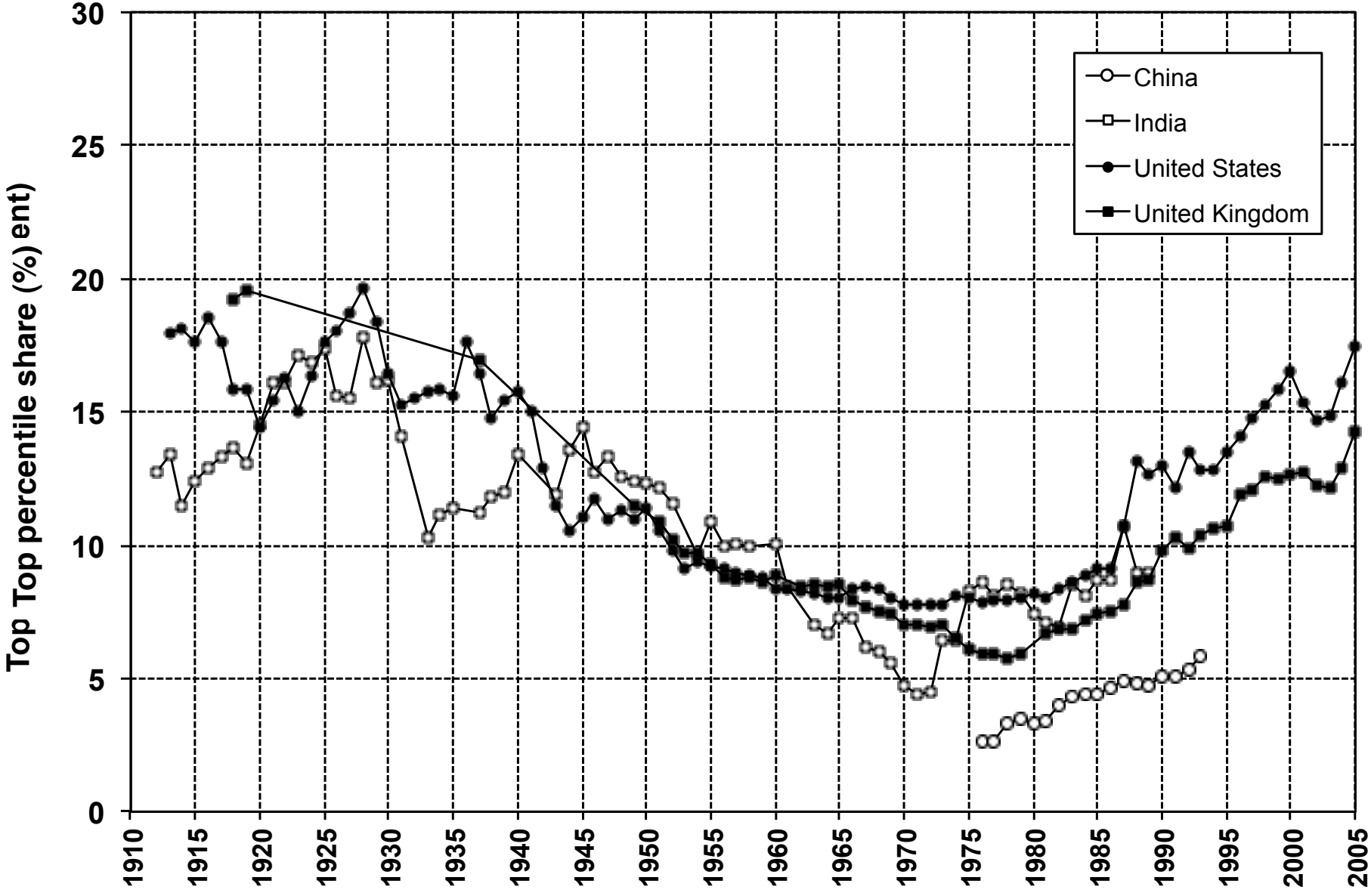
Consider **income inequality** in **China** and the **United States**:

> *"Over the last 30 years, top income shares have increased substantially in English speaking countries and in India and China but not in continental Europe countries or Japan."*
>
> *– Tony Atkinson, Thomas Piketty and Emmanuel Saez, "Top Incomes in the Long Run of History," 2010.*

> *"1 percent of the people take nearly a quarter of [U.S.] income."*
>
> *– Joseph Stiglitz, "Of the 1%, by the 1%, for the 1%," 2011.*

- The "top p%" principle uses percentiles to divide a distribution in 100 groups P1–P100 containing each 1% of its values.
- Common divisions are quartiles (Q1–Q4) containing each 25% of values, and deciles (D1–D10) containing each 10% of values.

Wealth concentration in the top 1% income share, 1920–2005



Source: Atkinson, Piketty and Saez (2010). Adapted from Fig. 7A and 7D.

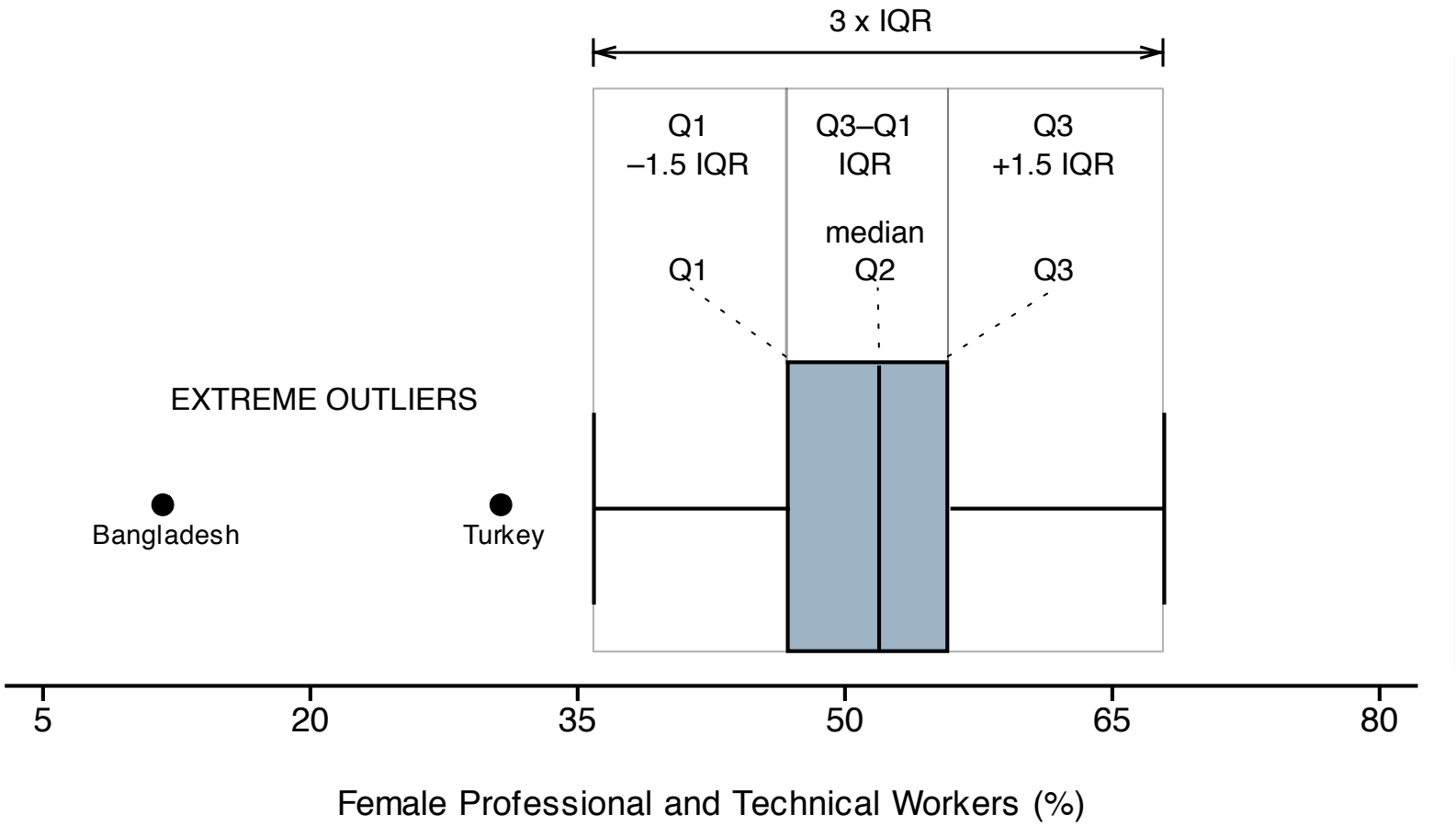
Quartiles, medians and box plots

Histograms are useful to compare a distribution with the normal distribution. Box plots have other purposes:

- **Detect outliers**, i.e. extreme observations:
  - Extreme observations distort the mean.
    e.g. Effect of Scandinavian countries on average female ministerial representation in democracies vs. dictatorships.
  - Extreme observations reduce normality.
    e.g. Effect of Turkey and especially Bangladesh on average and distribution of female worker rates in democracies.
- **Compare groups**, i.e. over categories:
  - With binary variables, e.g. male/female, democratic/dictatorial.
  - With nominal variables, e.g. religion, ethnicity, geography.
  - With interval variables, e.g. income groups, education levels.

Quartiles, medians and box plots

Box plot construction rules vary, but always show the median and 50% of the distribution as a 'box' with 'whiskers' at Q1 and Q3:



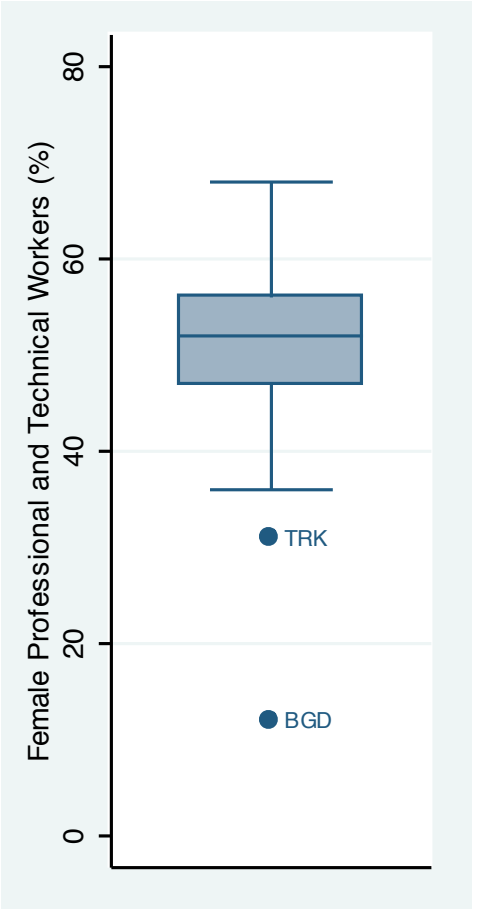
Stata: `gr box`

Box plot for worldwide female worker rates:

```
gr box workers, mark(1, mlabel(ccode))
```

- `gr box` produces vertical box plots, `gr hbox` produces horizontal ones. Select the best.
- `mark(1, mlabel(ccode))` labels outliers with the variable `ccode` (country codes here).
- To compare one or more variables across groups, use either `over` or `by`.

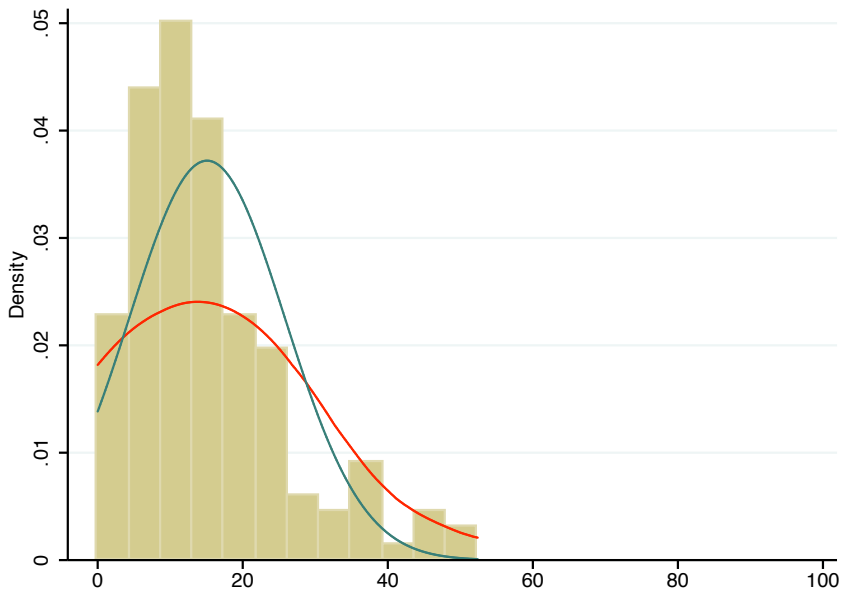
More examples appear in the course do-files and in the Stata Guide, Section 9.



Distributions

- μ stands for the mean (central tendency).
- σ^2 stands for variance, σ for the standard deviation (variability).

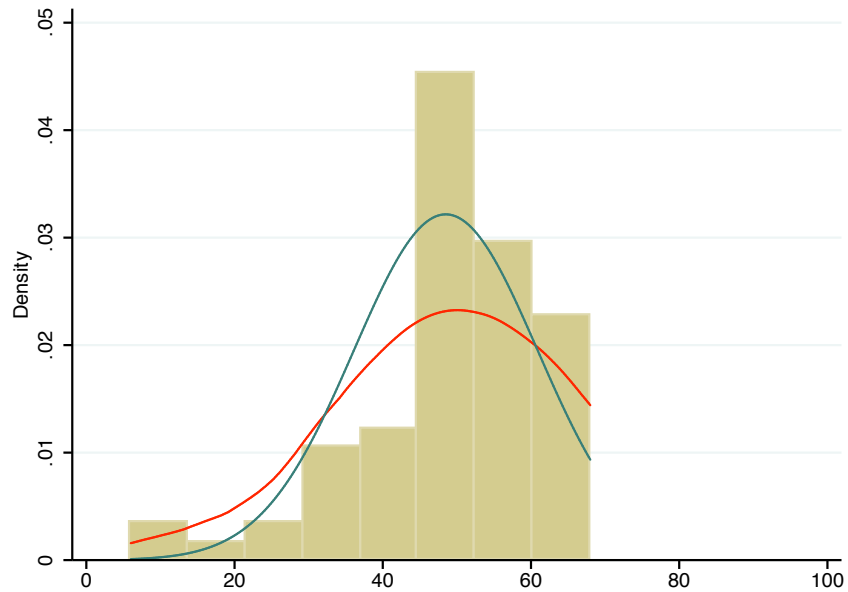
Female ministers (%)



$\mu = 15.08 \quad \sigma^2 = 115.03$

$N = 151 \quad \sigma = 10.72$

Female workers (%)

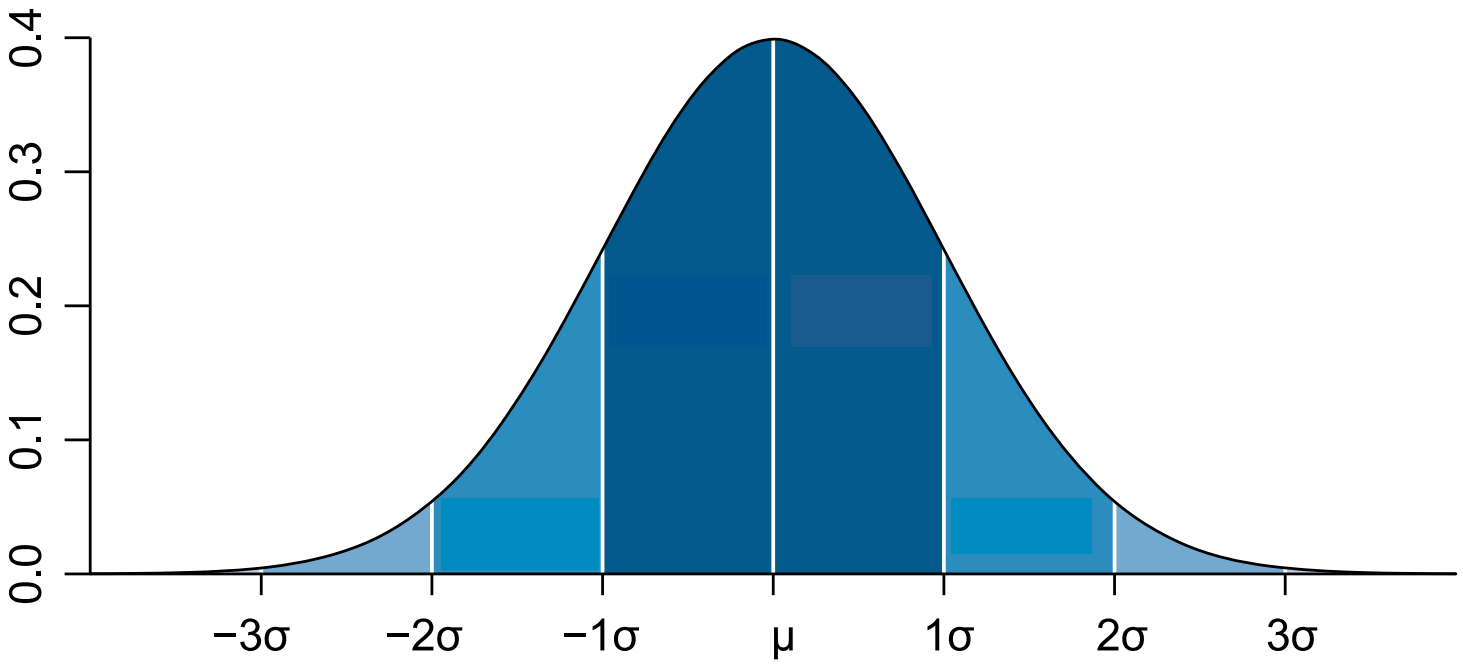


$\mu = 48.5 \quad \sigma^2 = 153.81$

$N = 74 \quad \sigma = 12.40$

Normal distribution $\mathcal{N}(\mu, \sigma^2)$

- In the standard normal distribution $\mathcal{N}(0, 1)$, $\mu = 0$ and $\sigma^2 = \sigma = 1$ (identical variance and standard deviation).
- In any normal distribution $\mathcal{N}(\mu, \sigma^2)$, all measures of central tendency are equal (identical mean, median and mode).



Normality

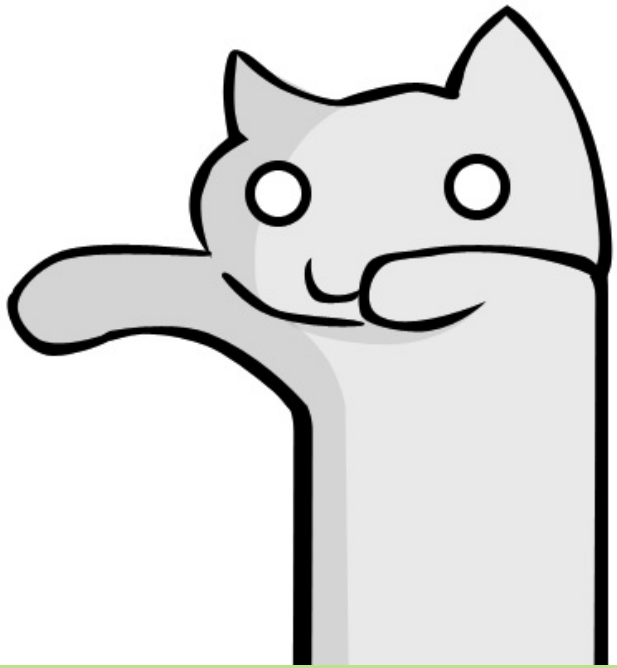
Normality assesses whether the distribution of a variable X approximates the normal distribution $\mathcal{N}$, written as $X \sim \mathcal{N}(\mu, \sigma^2)$.

Its most important properties to assess are:

- **Symmetry** around the mean/median/mode, i.e. null skewness.
- **Peakedness** and 'normal' tail sizes, i.e. 'normal' kurtosis.
- **Unimodality**, i.e. a single mode.

Note that the normal distribution is a theoretical construct that is *systematically violated* by the distributions of the data.

Violation of normality is acceptable, but only to *some* extent, given that estimation assumes a normal distribution of the data.



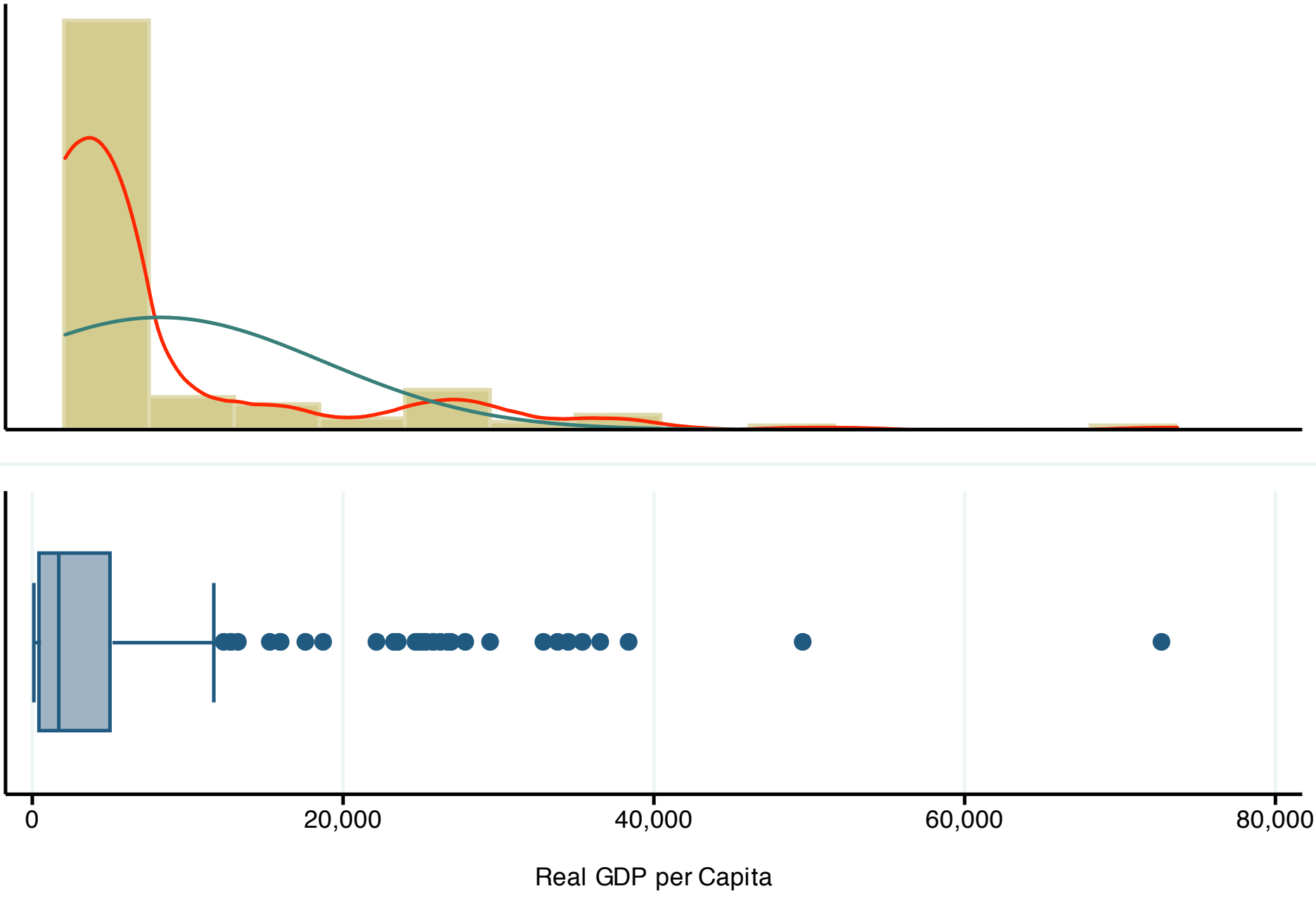
Stata: `hist`, `kdensity` etc.

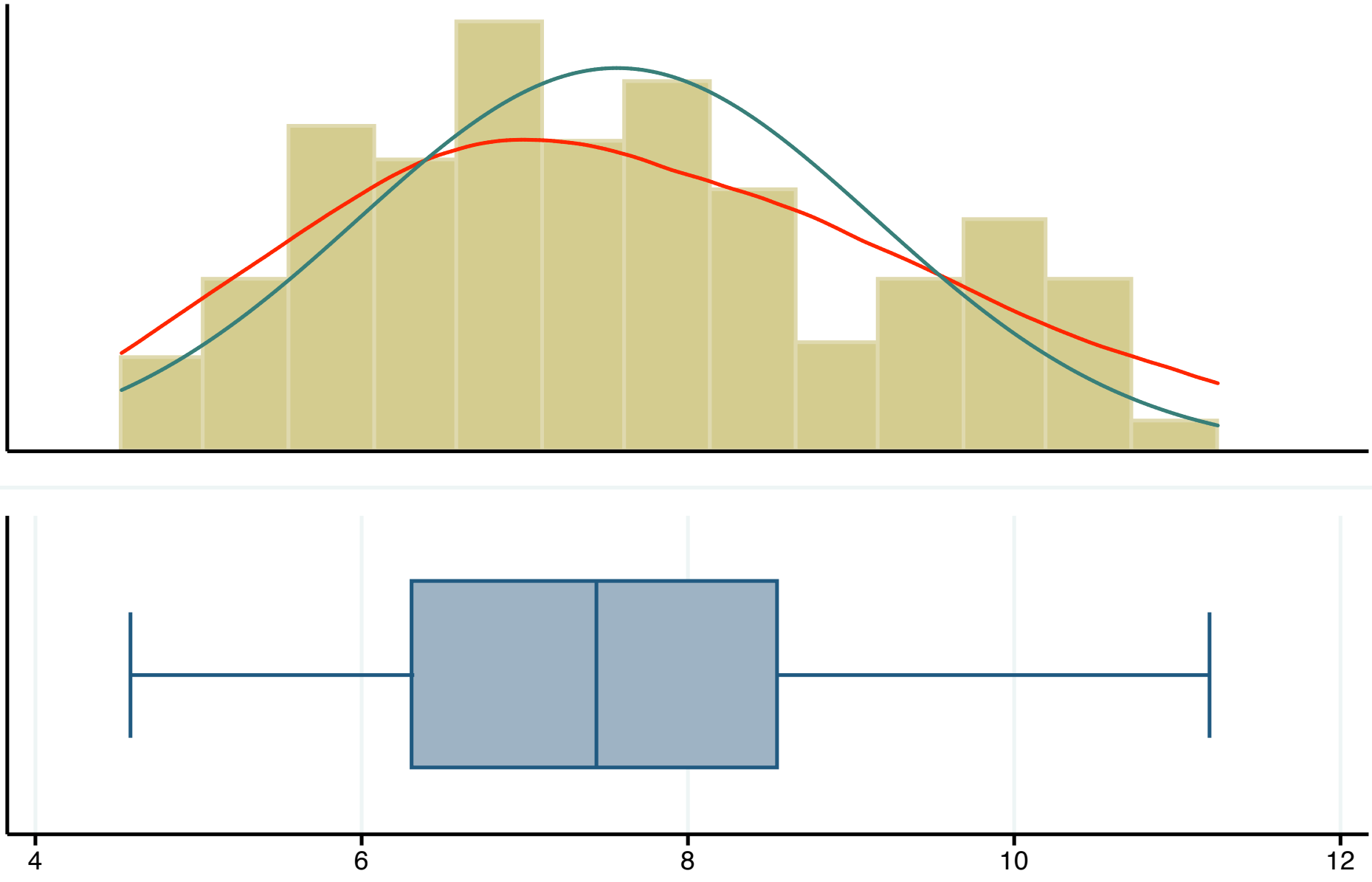
Visual assessment:

- Use `hist` (`histogram`) with the `normal` option to assess normality.
- Use `kdensity` to fit a kernel density curve to the distribution.
- Use `gr hbox` (horizontal box plot) to detect outliers.
- Use `symplot`, `pnorm` and `qnorm` for distributional diagnostic plots.

Formal assessment:

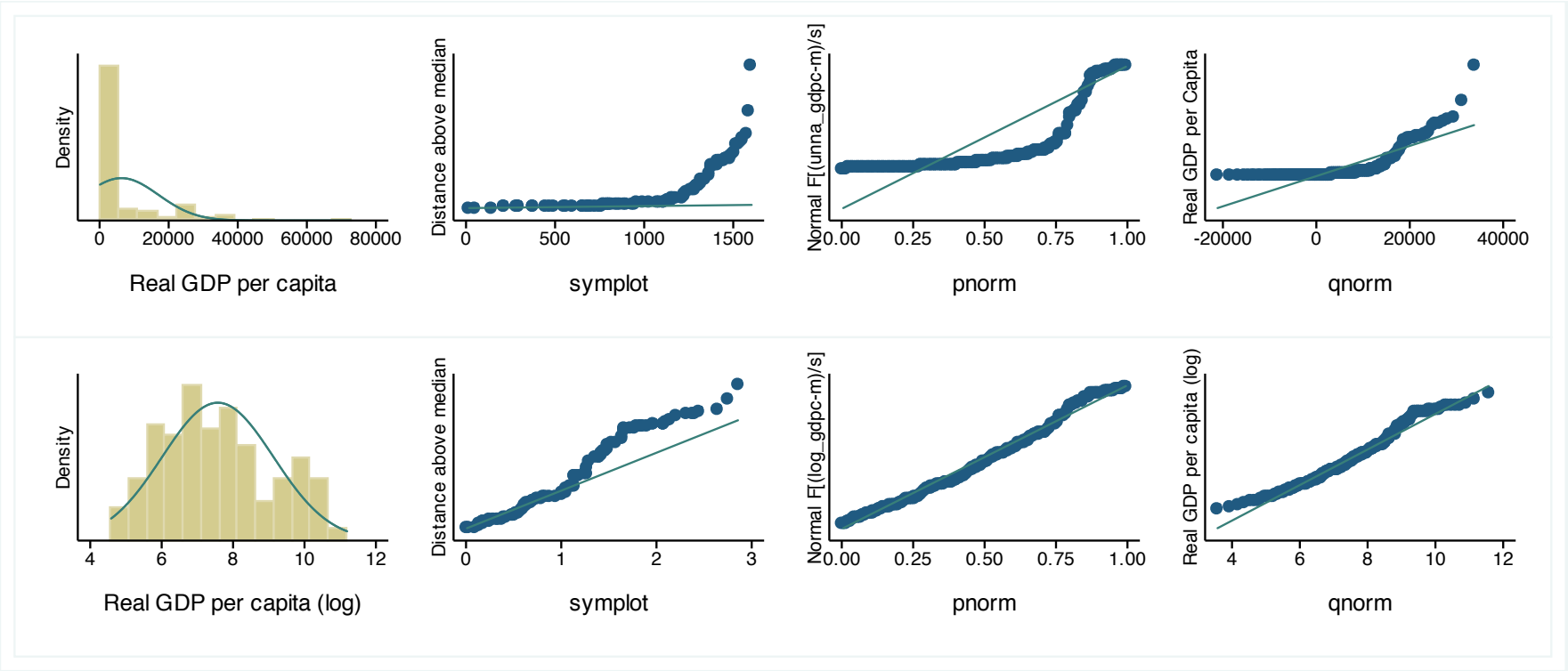
- Use `su` (`summarize`) with the `d` (`detail`) option to calculate whether *skewness* ≈ 0 and *kurtosis* ≈ 3.
- Use `tabstat` with the `iqr` (interquartile range) setting to calculate which observations are mild or extreme outliers.
- Use `ladder` to identify a possible transformation to a unit closer to normal distribution (usually to squared or log-units).


0
20,000
40,000
60,000
80,000
Real GDP per Capita


4
6
8
10
12
Real GDP per capita (log)

Stata implementation

Transform a variable only with a valuable method to reinterpret its new unit, as with exponentials: "log-GDP", "log-population" or with indices: "distance (inverse)," "quantity (squared)," etc.



Next time, The Prophecy.