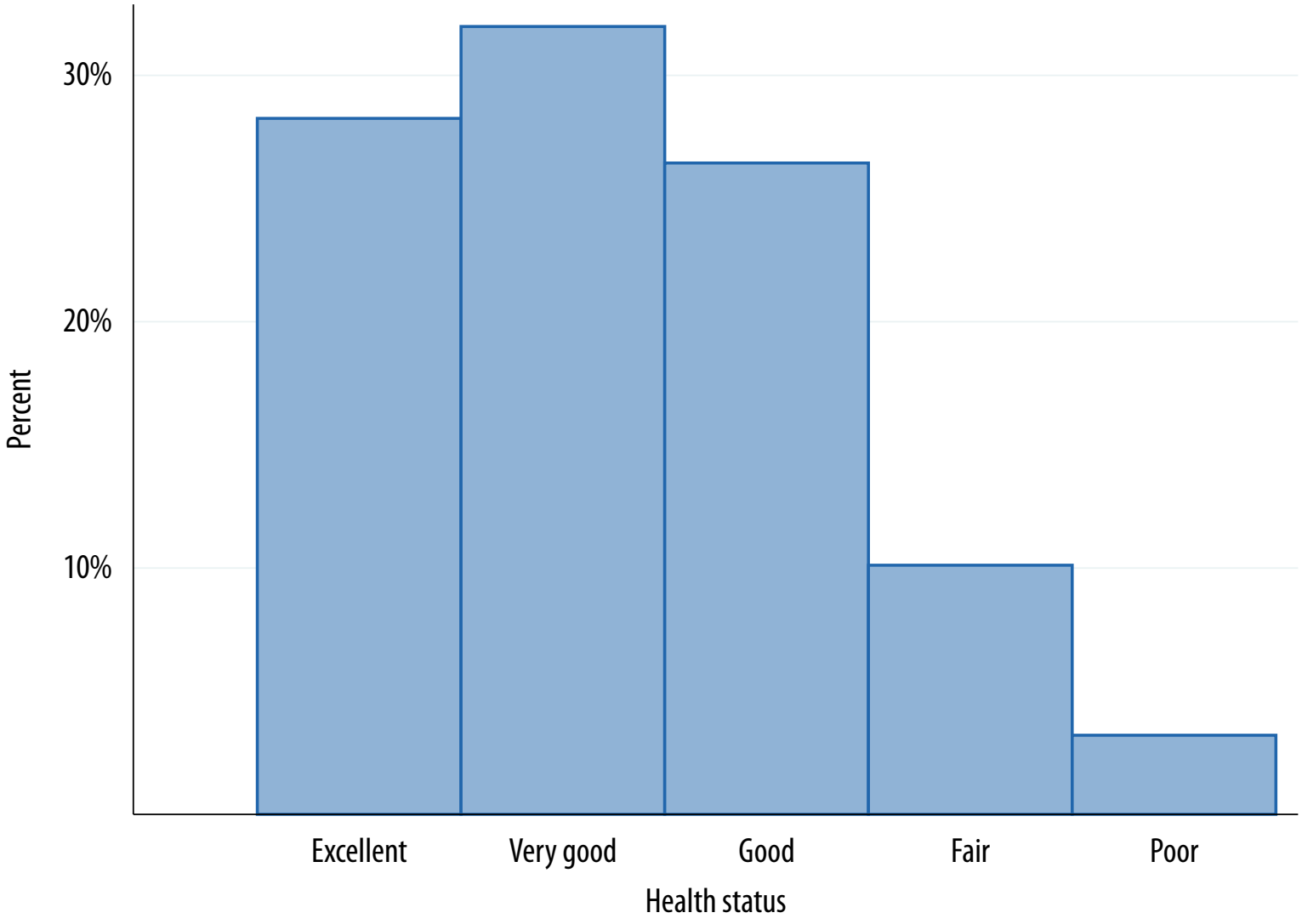
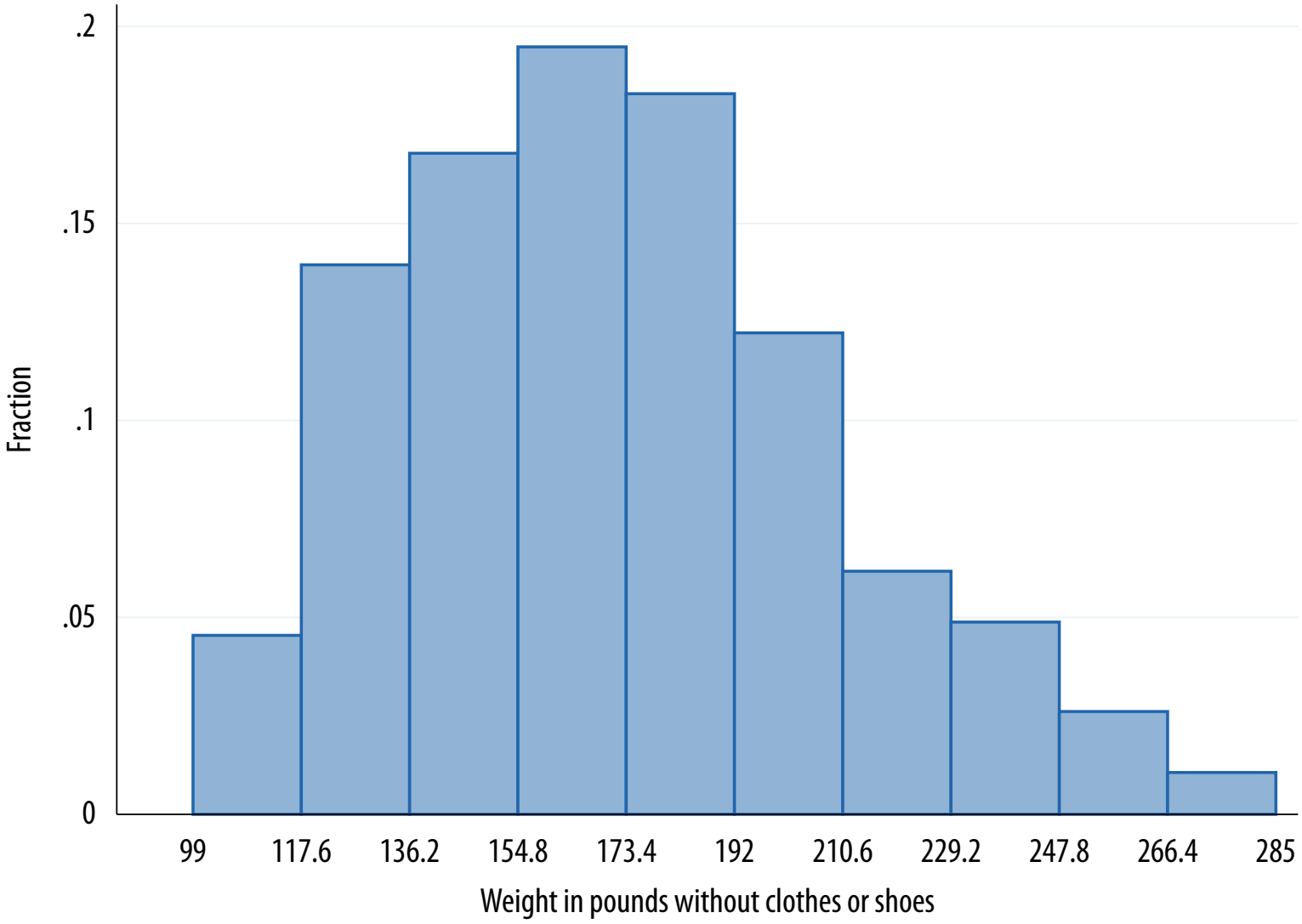
DISTRIBUTIONS

1. Histograms
2. Descriptors
3. Normal distribution
4. Practice

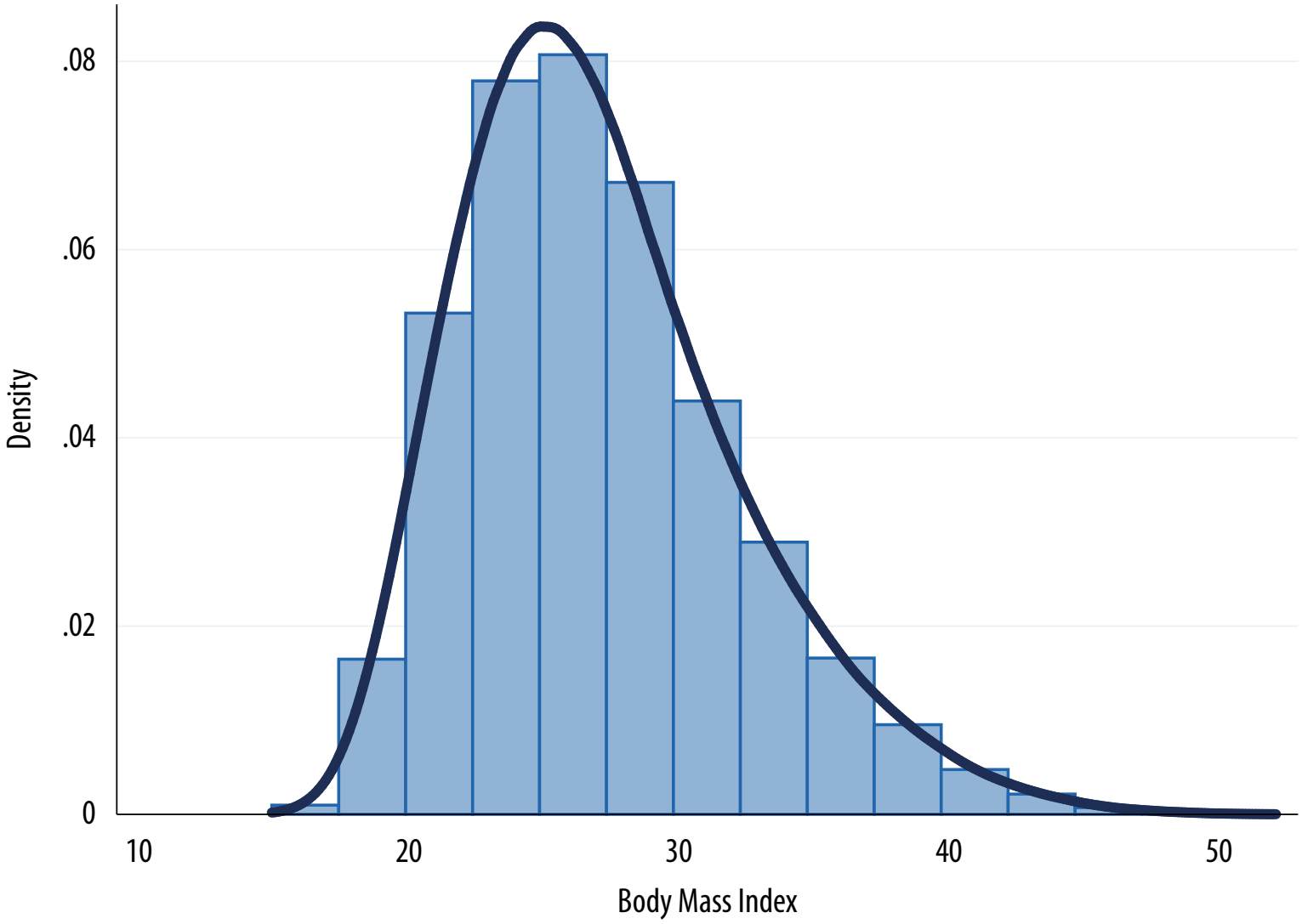
Frequencies of a categorical variable



Fractions of a continuous variable



Distribution of a variable



Histograms

Plot the distributions of continuous variables

- Use **histograms** for distributions `hist`
- Use **options** for better results `hist, bin(10) norm`
- Use **box plots** when there are outliers `gr (h)box`

Split distributions with categorical variables

- Plot **by** categories `hist, by(...)`
- Plot **over** categories `gr (h)box, over(...)`
- Plot **bars** to compare means `gr (h)bar, over(...)`

Descriptors

Measures of central tendency

- **Mean:** the ‘average’ value
- **Median:** the ‘middle’ value
- **Mode:** the ‘most frequent’ value

Usage

- Use the **mean** for continuous variables `su`
- Use the **median** when there are outliers `su, d`
- Use the **mode** for categorical variables `fre`

Mean and median

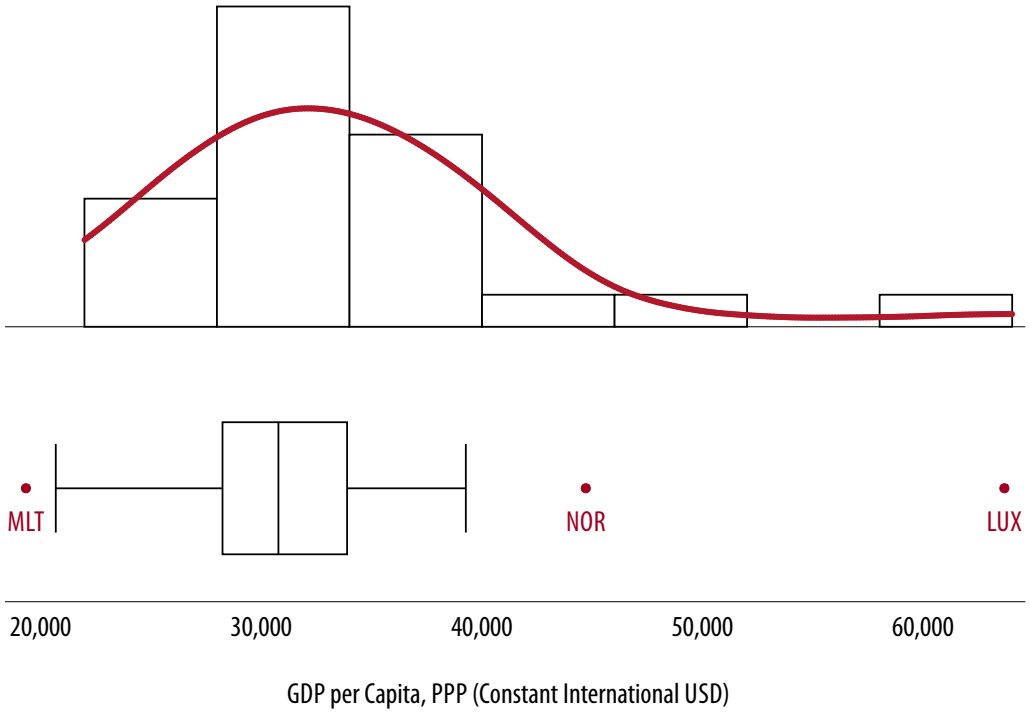
Arithmetic mean

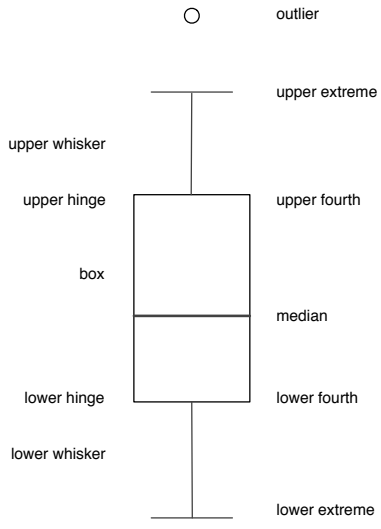
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Median value

- **Quartiles:** four segments each containing 25% of the data
- **Percentiles:** 100 segments each containing 1% of the data
- **Median:** 50th percentile (upper bound of ‘Q2’)

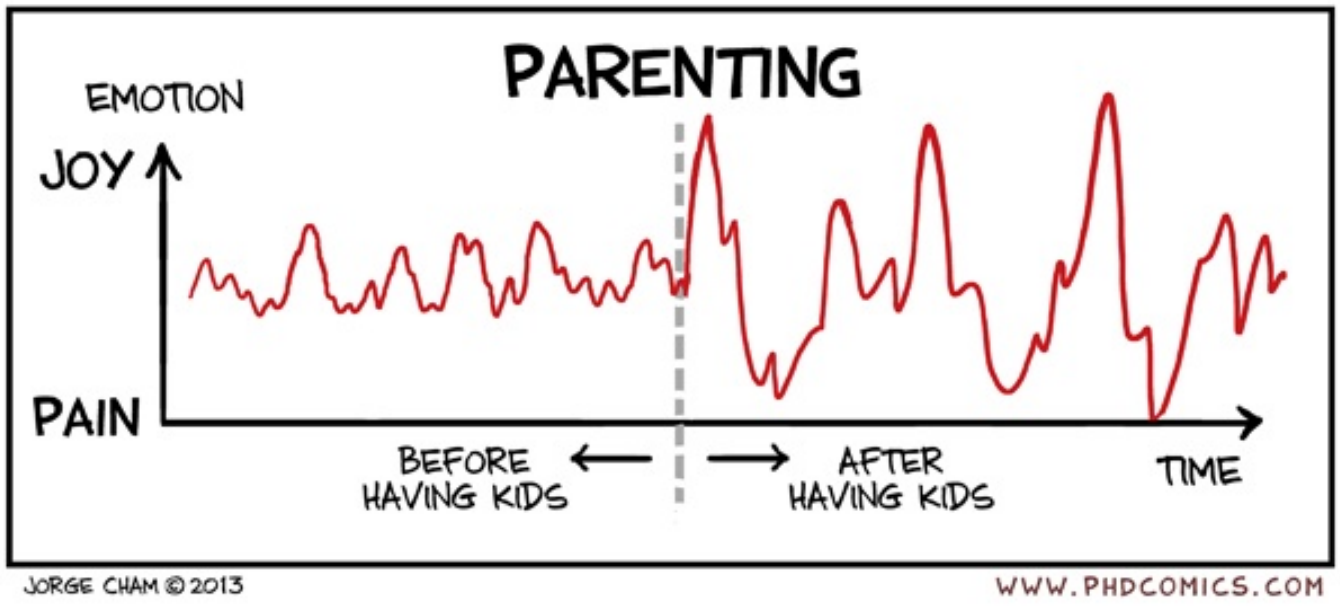
Skewness a.k.a. symmetry





- **Positive skew:** 'right tail' of higher values, mean $>$ median
- **Negative skew:** 'left tail' of lower values, mean $<$ median

Variability



Variability

Measures of dispersion

- $X_{max} - X_{min}$: **range** (in 'natural' units of X)
- Var_X: **squared distances from the mean** (summed over N)
- SD_X: **dispersion** (compound square root of variance by N)

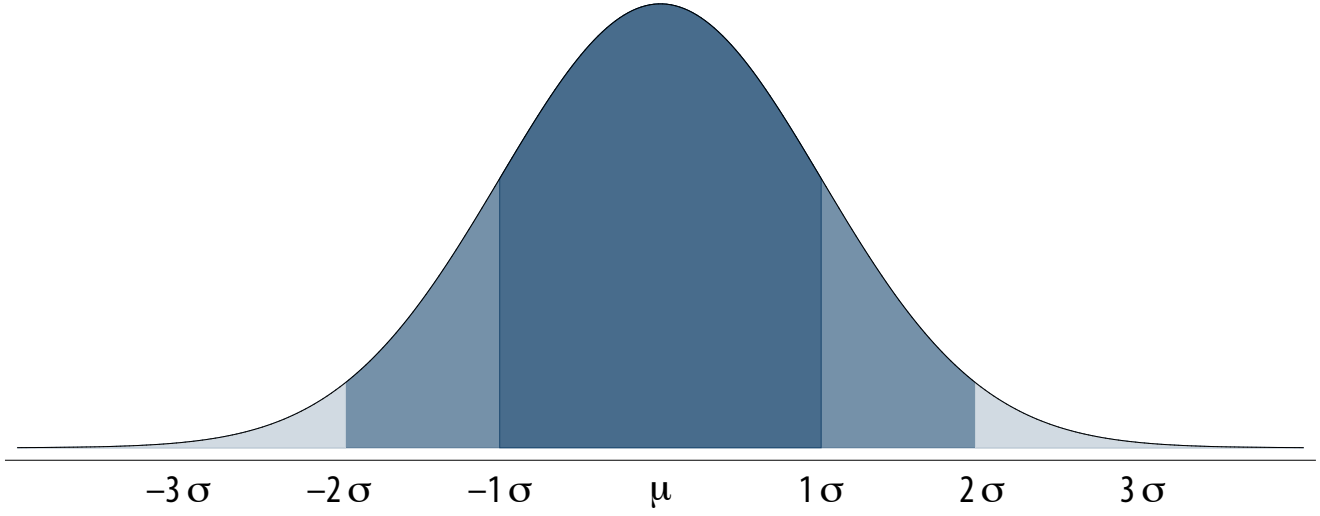
Variance and standard deviation

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

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

Properties

- $\mathcal{N}(\mu, \sigma^2)$: symmetric and unimodal
- $\mathcal{N}(\mu, \sigma^2)$: mean $=$ median $=$ mode
- $\mathcal{N}(0, 1)$: standard normal distribution



Normality assessment

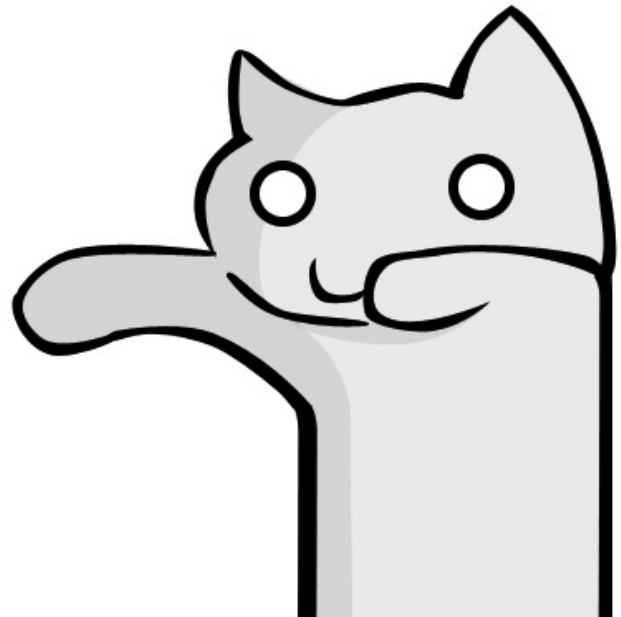
Visual assessment

- Distributions — `hist, normal`, `kdensity`, `gr (h)box`
- Diagnostics — `symplot`, `qnorm`, `(g)ladder`

Formal assessment

- Use `su x, d` to assess the symmetry (*skewness* ~ 0) and 'peakedness' (*kurtosis* ~ 3) of a variable.
- Use `tabstat x y, s(skew kurt) c(s)` to compare a variable with its transformation (often to log-units).

Next time, The Prophecy.



Practice: NHIS dataset

$$\text{Body Mass Index} = \frac{\text{mass (kg)}}{(\text{height(m)})^2} = \frac{\text{mass (lb)} \times 703}{(\text{height(in)})^2}$$

- For **normal weight** adults, $18.5 < \text{BMI} < 25$.
- For **overweight** adults, $25 \leq \text{BMI} < 30$.
- For **obese** adults, $\text{BMI} \geq 30$.

Data:

- National Health Interview Survey (NHIS)
- Sample: U.S. adult population, 2009

Practice session

Class

```
* Get the do-file for this week.
srqm fetch week4.do
* Open to read and replicate.
doedit code/week4
```

Coursework

- Finish the do-file and read all comments at home.
- Follow instructions on top of the code.
- Prepare questions in your group's draft do-file.

Exercise

Ex 4.1. Quality of Government 2011

1. What countries have *much* more females in government?
2. How is the female-to-male income ratio distributed?
3. Same question with confidence variables (`d wvs_e069*`).
4. Plot the Gini coefficient over quartiles of GDP per capita.

Tips

- Label outliers: `gr hbox x, mark(1, mlab(ccodealp))`
- Get quartiles: `xtile qx = x, nq(4)`