Unit 3: Descriptive Statistics for Continuous Data

Statistics for Linguists with R – A SIGIL Course

Designed by Marco Baroni[1] and Stefan Evert[2]

[1]Center for Mind/Brain Sciences (CIMeC)
University of Trento, Italy

[2]Institute of Cognitive Science (IKW)
University of Osnabrück, Germany

Outline

Outline

Reminder: the library metaphor

- In the library metaphor, we took random samples from an infinite population of tokens (words, VPs, sentences, ...)
- Relevant property is a binary (or **categorical**) classification
  - active vs. passive VP or sentence (binary)
  - instance of lemma TIME vs. some other word (binary)
  - subcategorisation frame of verb token (itr, tr, ditr, p-obj, ...)
  - part-of-speech tag of word token (50+ categories)
- Characterisation of population distribution is straightforward
  - **binomial**: true proportion $\pi = 10\%$ of passive VPs, or relative frequency of TIME, e.g. $\pi = 2000$ pmw
  - alternatively: specify redundant proportions $(\pi, 1 - \pi)$, e.g. passive/active VPs $(.1, .9)$ or TIME/other $(.002, .998)$
  - **multinomial**: multiple proportions $\pi_1 + \pi_2 + \cdots + \pi_K = 1$, e.g. $(\pi_{\text{noun}} = .28, \pi_{\text{verb}} = .17, \pi_{\text{adj}} = .08, \ldots)$

Numerical properties

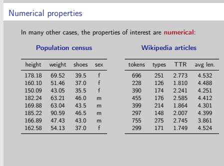
In many other cases, the properties of interest are **numerical**:

Population census

height	weight	shoes	sex
178.18	69.52	39.5	f
160.10	51.46	37.0	f
150.09	43.05	35.5	f
182.24	63.21	46.0	m
169.88	63.04	43.5	m
185.22	90.59	46.5	m
166.89	47.43	43.0	m
162.58	54.13	37.0	f

Wikipedia articles

tokens	types	TTR	avg len.
696	251	2.773	4.532
228	126	1.810	4.488
390	174	2.241	4.251
455	176	2.585	4.412
399	214	1.864	4.301
297	148	2.007	4.399
755	275	2.745	3.861
299	171	1.749	4.524



1. Traditional example: populace of country, i.e. population of all inhabitants. Properties of interest are physical measurements such as height, weight and shoe size; also age, income, IQ, size of household, . . .
2. A more linguistic example: population of all English Wikipedia articles, with frequency statistics such as token count, type count, proportion of passives, token-type-ratio (TTR), avg. word length (wrt. tokens or types), avg. frequency/familiarity class, . . .
3. NB: both populations are finite, but very large ("practically infinite").
4. Often there are also categorical properties, e.g. sex, level of education, Wikipedia category, has page won an award?, . . .

Descriptive vs. inferential statistics

Two main tasks of "classical" statistical methods (numerical data):

1. **Descriptive statistics**
   - compact description of the distribution of a (numerical) property in a very large or infinite population
   - often by characteristic **parameters** such as mean, variance, . . .
   - this was the original purpose of statistics in the 19th century
2. **Inferential statistics**
   - infer (aspects of) population distribution from a comparatively small random sample
   - accurate estimates for level of uncertainty involved
   - often by testing (and rejecting) some **null hypothesis** H_0

Outline

Statisticians distinguish 4 scales of measurement

Categorical data

1. **Nominal scale**: purely qualitative classification
   - male vs. female, passive vs. active, POS tags, subcat frames
2. **Ordinal scale**: ordered categories
   - school grades A–E, social class, low/medium/high rating

Numerical data

3. **Interval scale**: meaningful comparison of differences
   - temperature (°C), plausibility & grammaticality ratings
4. **Ratio scale**: comparison of magnitudes, absolute zero
   - time, length/width/height, weight, frequency counts

Additional dimension: **discrete** vs. **continuous** numerical data

- discrete: frequency counts, rating $(1, \dots, 7)$, shoe size, …
- continuous: length, time, weight, temperature, …

Quiz

Which scale of measurement / data type is it?

- subcategorisation frame
- reaction time (in psycholinguistic experiment)
- familiarity rating on scale $1, \dots, 7$
- room number
- grammaticality rating: "*", "??", "?" or "ok"
- magnitude estimation of plausibility (graphical scale)
- frequency of passive VPs in text
- relative frequency of passive VPs
- token-type-ratio (TTR) and average word length (Wikipedia)

☞ in this unit: continuous numerical variables on ratio scale

Outline

The task

- Census data from small country of *Ingary* with $m = 502{,}202$ inhabitants. The following properties were recorded:
  - body height in cm
  - weight in kg
  - shoe size in Paris points (Continental European system)
  - sex (*male*, *female*)
- Frequency statistics for $m = 1{,}429{,}649$ Wikipedia articles:
  - token count
  - type count
  - token-type ratio (TTR)
  - average word length (across tokens)

☞ Describe / summarise these data sets (continuous variables)

```
> library(corpora)
> FakeCensus <- simulated.census()
> WackypediaStats <- simulated.wikipedia()
```

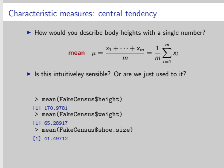
Characteristic measures: central tendency

- How would you describe body heights with a single number?

$$\textbf{mean} \quad \mu = \frac{x_1 + \dots + x_m}{m} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Is this intuitiveley sensible? Or are we just used to it?

```
> mean(FakeCensus$height)
[1] 170.9781
> mean(FakeCensus$weight)
[1] 65.28917
> mean(FakeCensus$shoe.size)
[1] 41.49712
```

1. We will see a (partial) mathematical justification later today.

Characteristic measures: variability (spread)

- Average weight of 65.3 kg not very useful if we have to design an elevator for 10 persons or a chair that doesn't collapse: We need to know if everyone weighs close to 65 kg, or whether the typical range is 40–100 kg, or whether it is even larger.
- Measure of spread: **minimum** and **maximum**, here 30–196 kg
- We're more interested in the "typical" range of values without the most extreme cases
- Average variability based on **error** $x_i - \mu$ for each individual shows how well the mean μ describes the entire population

$$\textbf{variance} \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$$

Characteristic measures: variability (spread)

$$\textbf{variance} \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$$

☞ Do you remember how to calculate this in R?
  - height: $\mu = 171.00$, $\sigma^2 = 199.50$, $\sigma = 14.12$
  - weight: $\mu = 65.29$, $\sigma^2 = 306.72$, $\sigma = 17.51$
  - shoe size: $\mu = 41.50$, $\sigma^2 = 21.70$, $\sigma = 4.66$

- Mean and variance are not on a comparable scale
  ➜ **standard deviation** (s.d.) $\sigma = \sqrt{\sigma^2}$
- NB: still gives more weight to larger errors!

Characteristic measures: higher moments

- Mean based on $(x_i)^1$ is also known as a "first moment", variance based on $(x_i)^2$ as a "second moment"
- The third moment is called **skewness**

$$\gamma_1 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_i-\mu}{\sigma}\right)^3$$

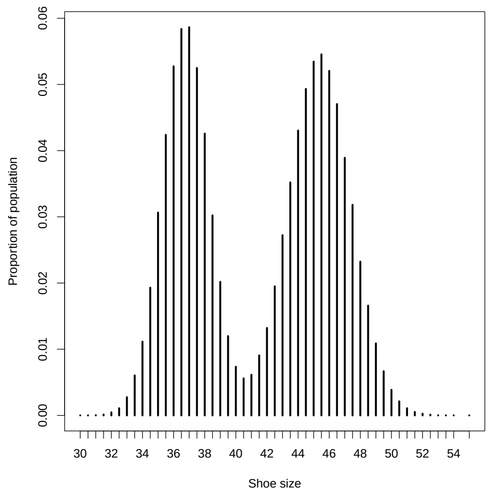
  and measures the asymmetry of a distribution
- The fourth moment (**kurtosis**) measures "bulginess"

- How useful are these characteristic measures?
  - Given the mean, s.d., skewness, ..., can you tell how many people are taller than 190 cm, or how many weigh $\approx$ 100 kg?
  - Such measures mainly used for computational efficiency, and even this required an elaborate procedure in the 19th century

Outline

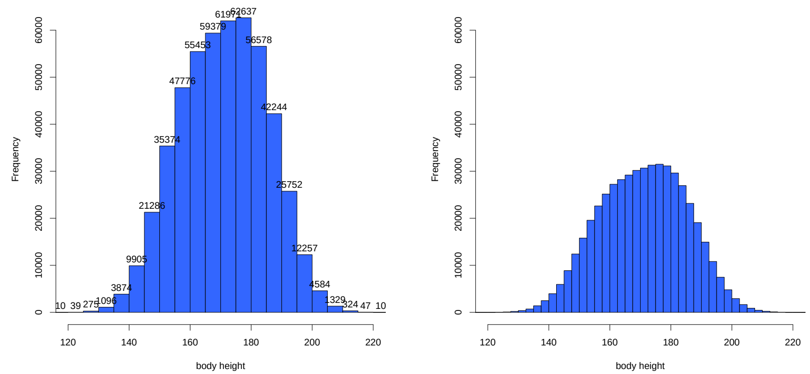
The shape of a distribution: discrete data

Discrete numerical data can be tabulated and plotted



The shape of a distribution: histogram for continuous data

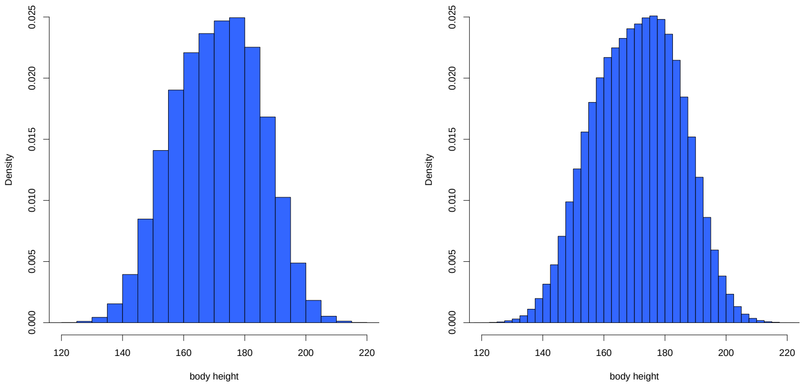
Continuous data must be collected into bins ➜ histogram



- No two people have *exactly* the same body height, weight, ...
- Frequency counts (= y-axis scale) depend on number of bins

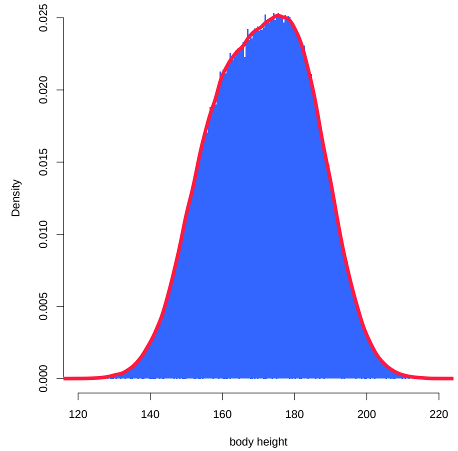
The shape of a distribution: histogram for continuous data

Continuous data must be collected into bins ➜ histogram



- **Density** scale is comparable for different numbers of bins
- Area of histogram bar $\equiv$ relative frequency in population

Refining histograms: the density function



- Contour of histogram = **density function**

Outline

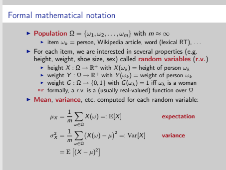
Formal mathematical notation

- **Population** $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ with $m \approx \infty$
  - item ω_k = person, Wikipedia article, word (lexical RT), …
- For each item, we are interested in several properties (e.g. height, weight, shoe size, sex) called **random variables** (**r.v.**)
  - height $X : \Omega \to \mathbb{R}^+$ with $X(\omega_k)$ = height of person ω_k
  - weight $Y : \Omega \to \mathbb{R}^+$ with $Y(\omega_k)$ = weight of person ω_k
  - weight $G : \Omega \to \{0, 1\}$ with $G(\omega_k) = 1$ iff ω_k is a woman
  - ☞ formally, a r.v. is a (usually real-valued) function over Ω
- **Mean**, **variance**, etc. computed for each random variable:

$$\mu_X = \frac{1}{m} \sum_{\omega \in \Omega} X(\omega) =: \mathrm{E}[X] \qquad \textbf{expectation}$$

$$\sigma_X^2 = \frac{1}{m} \sum_{\omega \in \Omega} \bigl(X(\omega) - \mu\bigr)^2 =: \mathrm{Var}[X] \qquad \textbf{variance}$$

$$= \mathrm{E}\left[(X - \mu)^2\right]$$



1. Term *random variable* makes sense if you think of e.g. X as the height of a randomly selected person; it yields a different value each time you pick a new person.
2. Point out the numerical $\{0, 1\}$ coding of a binary categorical variable. If numerical coding is used for multinomial variables, each category hast to be coded by a separate indicator variable; otherwise, spurious relations between the categories would be implied.
3. Keep in mind that $\mu \in \mathbb{R}$ is a fixed real number, so the variance sum can be calculated as an expectation over (a function of) X.

Working with random variables

- $X'(\omega) := \bigl(X(\omega) - \mu\bigr)^2$ defines new r.v. $X' : \Omega \to \mathbb{R}$
  - ☞ any function $f(X)$ of a r.v. is itself a random variable
- The expectation is a **linear functional** on r.v.:
  - $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ for $X, Y : \Omega \to \mathbb{R}$
  - $\mathrm{E}[r \cdot X] = r \cdot \mathrm{E}[X]$ for $r \in \mathbb{R}$
  - $\mathrm{E}[a] = a$ for constant r.v. $a \in \mathbb{R}$ (additional property)
- These rules enable us to simplify the computation of σ_X^2:

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X - \mu_X)^2\right] = \mathrm{E}\left[X^2 - 2\mu_X X + \mu_X^2\right]$$
$$= \mathrm{E}[X^2] - 2\mu_X \underbrace{\mathrm{E}[X]}_{=\mu_X} + \mu_X^2 = \mathrm{E}[X^2] - \mu_X^2$$

- Random variables and probabilities: r.v. X describes outcome of picking a random $\omega \in \Omega$ ➜ **sampling distribution**

$$\Pr(a \le X \le b) = \frac{1}{m}\bigl|\{\omega \in \Omega \,|\, a \le X(\omega) \le b\}\bigr|$$

A justification for the mean

- σ_X^2 tells us how well the r.v. X is characterised by μ_X
- More generally, $\mathrm{E}\left[(X-a)^2\right]$ tells us how well X is characterised by some real number $a \in \mathbb{R}$
- The best single value we can give for X is the one that minimises the average squared error:

$$\mathrm{E}\left[(X-a)^2\right] = \mathrm{E}[X^2] - 2a\underbrace{\mathrm{E}[X]}_{=\mu_X} + a^2$$

- It is easy to see that a minimum is achieved for $a = \mu_X$
  - ☞ The quadratic error term in our definition of σ_X^2 guarantees that there is always a unique minimum. This would not have been the case e.g. with $|X - a|$ instead of $(X-a)^2$.

How to compute the expectation of a discrete variable

- Population distribution of a **discrete** variable is fully described by giving the relative frequency of each possible value $t \in \mathbb{R}$:

$$\pi_t = \Pr(X = t)$$

$$\mathrm{E}[X] = \sum_{\omega\in\Omega} \frac{X(\omega)}{m} = \underbrace{\sum_t \sum_{X(\omega)=t}}_{\text{group by value of } X} \frac{t}{m} = \sum_t t \sum_{X(\omega)=t} \frac{1}{m}$$
$$= \sum_t t \cdot \frac{|X(\omega) = t|}{m} = \sum_t t \cdot \pi_t = \sum_t t \cdot \Pr(X = t)$$

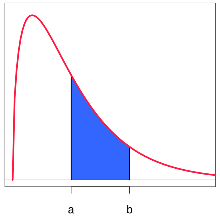
- The second moment $\mathrm{E}[X^2]$ needed for $\mathrm{Var}[X]$ can also be obtained in this way from the population distribution:

$$\mathrm{E}[X^2] = \sum_t t^2 \cdot \Pr(X = t)$$

How to compute the expectation of a continuous variable

- Population distribution of **continuous** variable can be described by its **density function** $g : \mathbb{R} \to [0, \infty]$
  - keep in mind that $\Pr(X = t) = 0$ for almost every value $t \in \mathbb{R}$: nobody is *exactly* 172.3456789 cm tall!

Area under density curve between a and b = proportion of items $\omega \in \Omega$ with $a \leq X(\omega) \leq b$.



$$\Pr(a \leq X \leq b) = \int_a^b g(t)\, dt$$

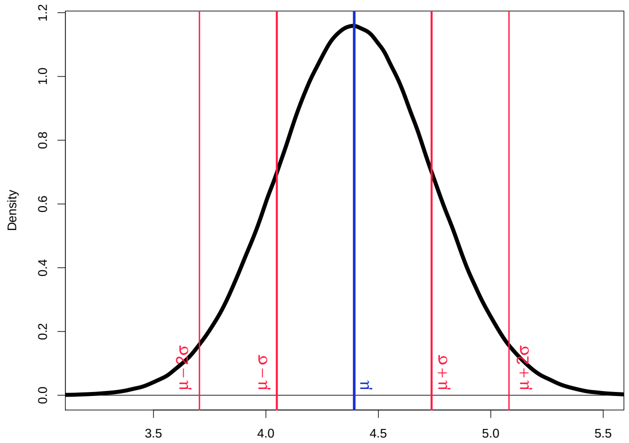
Same reasoning as for discrete variable leads to:

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} t \cdot g(t)\, dt \qquad \text{and}$$

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(t) \cdot g(t)\, dt$$

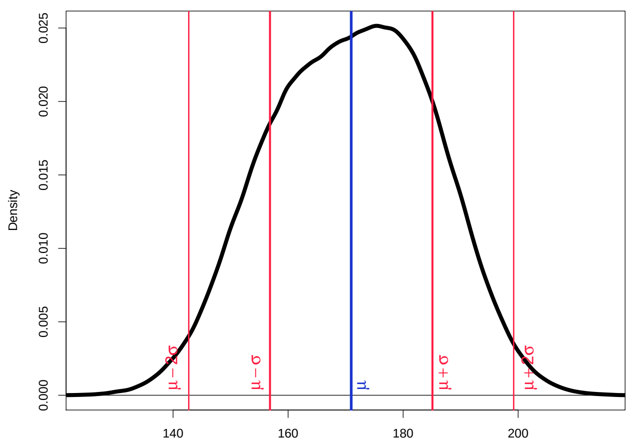
Outline

Different types of continuous distributions



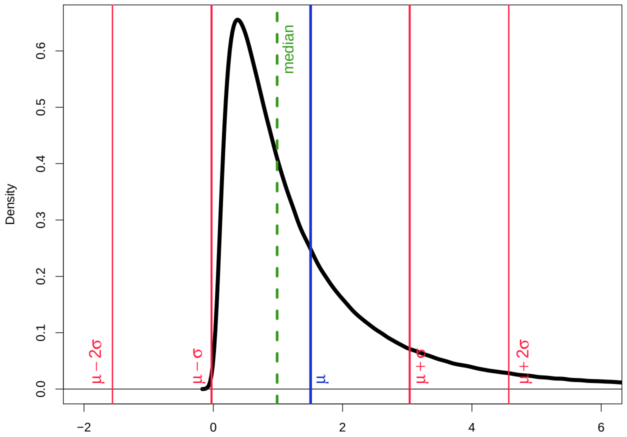
symmetric, bell-shaped

Different types of continuous distributions



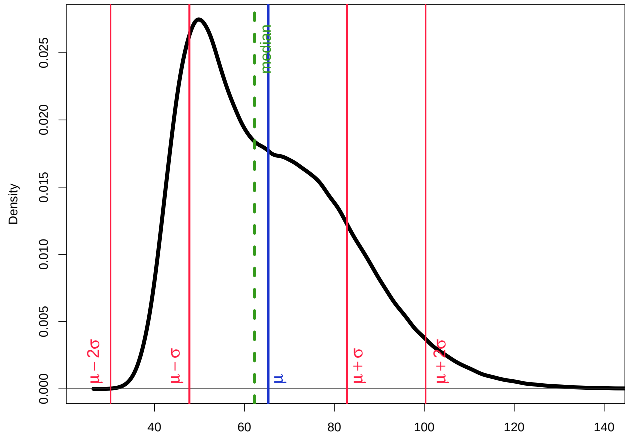
symmetric, bulgy

Different types of continuous distributions



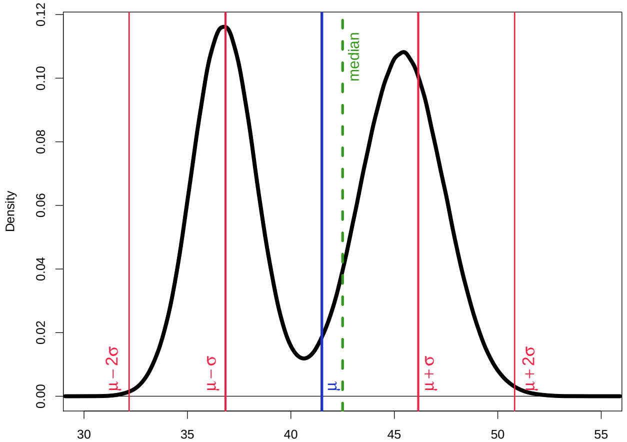
skewed (median $\neq$ mean)

Different types of continuous distributions

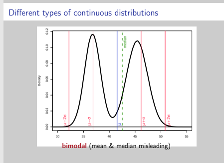


complicated . . .

Different types of continuous distributions



bimodal (mean & median misleading)



1. For each distribution type, ask participants how well the population is described by μ and σ.
2. Explain concepts of median and mode on these examples.
3. Optional task: look at example data and compute histograms. Which distribution types do you find?

Outline

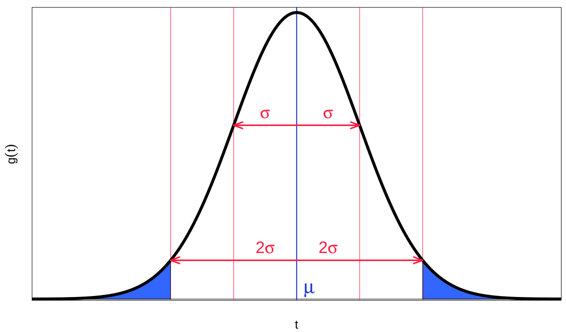
The Gaussian distribution

- In many real-life data sets, the distribution has a typical "bell-shaped" form known as a **Gaussian** (or **normal**)

- Idealised density function is given by simple equation:

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2}$$

with parameters $\mu \in \mathbb{R}$ (location) and $\sigma > 0$ (width)



- Notation: $X \sim N(\mu, \sigma^2)$ if r.v. has such a distribution
- No coincidence: $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$ (➜ homework ;-)

Important properties of the Gaussian distribution

- Distribution is well-behaved: symmetric, and most values are relatively close to the mean μ (within 2 standard deviations)

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2}\, dt \approx 95.5\%$$

  - 68.3% are within range $\mu - \sigma \leq X \leq \mu + \sigma$ (one s.d.)

- The **central limit theorem** explains why this particular distribution is so widespread (sum of independent effects)

☞ Mean and standard deviation are meaningful characteristics if distribution is Gaussian or near-Gaussian
  - completely determined by these parameters