3 Data collection & samples

AUTHOR
Thomas Brochhagen

AFFILIATION
Universitat Pompeu Fabra

PUBLISHED
April 5, 2024

Contents

Samples and their size

A sample is a collection of data points.

Your sample is most likely not the entire population you care about. That is, it is likely not *a complete sample*. Sample size is consequently almost always important, but in which way it matters is a function of your research question (chiefly: how large/small you expect the effect that you are trying to estimate to be). The truism that "more is better" applies but it does not follow that little may not already be enough.

A sample can be many things. It can be *unbiased* and *representative*; that is, sampled from the complete sample with a procedure that does not depend on the data points being sampled. But it can also be *biased* in some way. For instance, you may want to study a property of spoken Catalan but your sample may be comprised only of data from Catalan speakers from the Vallès Oriental between 20-30 years old. A priori, this is not a problem but it is something that needs to be factored in.

Analysis checks

It is always a good idea to check your analysis plan throughout all stages. There are two important ways to do so right after devising your plan. It is good practice to do both.

Pilot studies

A pilot study is a small-scale version of your entire analysis. It serves as a proof of concept, checking both the feasibility and quality of the analysis plan. It is also a safeguard to make sure you are not spending a lot of time and money on collecting a sample that may not be adequate for your question.

Simulation studies

A simulation study uses computational methods to generate virtual data that simulates the data you will collect. This allows you to think through your entire analysis and check how sample size (see below) and estimated effects may interact.

Where to get data

- Laboratory (e.g., eye-tracking or fMRI data)
- Online platforms (e.g., Amazon's Mechanical Turk or Prolific)
- Models (e.g., distributed meaning representations or Language Models)
- Corpora (e.g., Wikipedia, Twitter)
- Surveys (e.g., asking people in the "field")
- Available data sets, e.g.,
  - NoRaRe: A Multilingual Database of Word and Concept Properties
  - Small World of Words: Associativity data
  - SimLex-999: word similarity ratings
  - WordSim353: word relatedness ratings
  - WordNet: A lexical database
  - CLICS: Database of Cross-Linguistic Colexifications
  - ... and most modern (and responsible) studies involving data

Distributions

A probability distribution describes a random phenomenon (drawing a sample from it) in terms of (i) a sample space (possible values you can expect) and (ii) the probabilities of events (e.g., drawing a with a value greater than 3.4). Rephrasing the same thing more intuitively, a distribution defines all the possible values you can expect in an experiment and how likely they are.

Different distributions are useful to speak about different things. We will focus on three important ones: Normal/Gaussian; Binomial; and Poisson.

Normal distribution

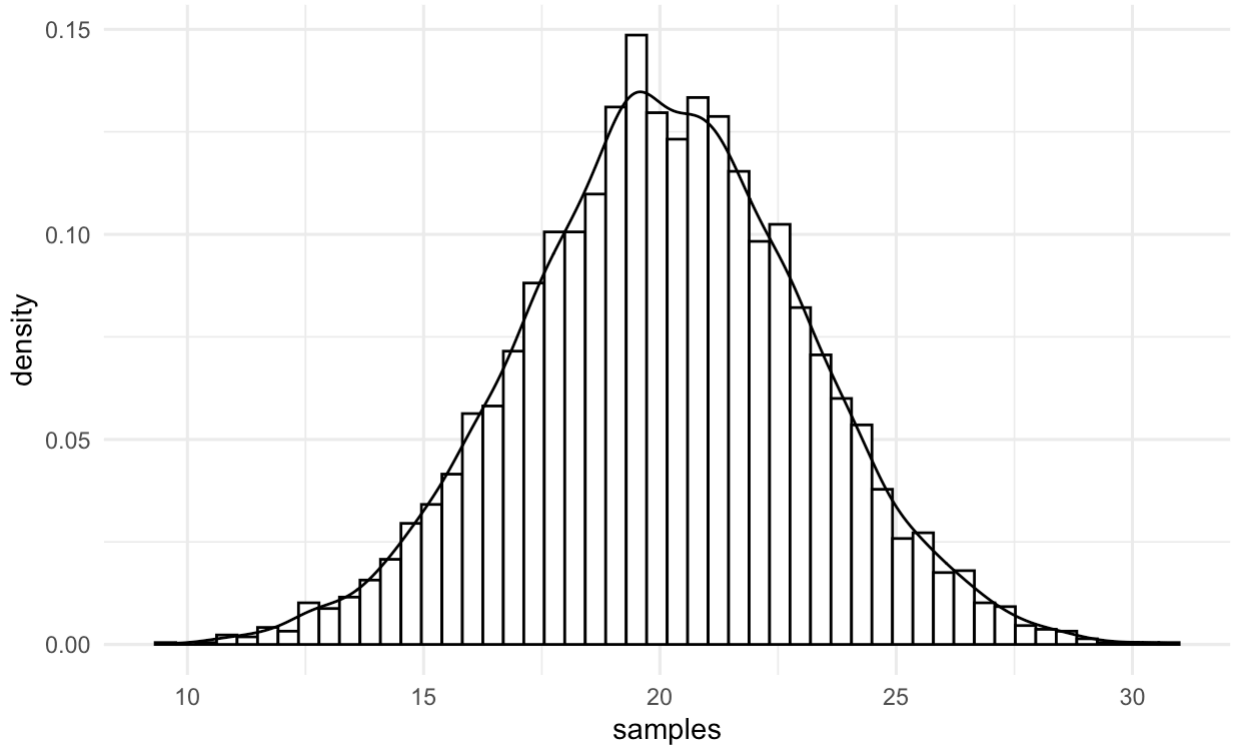
A Normal distribution is defined by a mean (often described by the letter μ) and a standard deviation (often described by σ). You can expect samples from a Normal distribution to center around the mean and to become less likely the further they are away from it. The normal distribution spans across real numbers.

Here's an illustration of 5000 samples drawn from a normal distribution with mean 20 and standard deviation of 3:

```
set.seed(       333     )
library (       ggplot2 )

samples  <-       rnorm   (       n =        5000    , mean =        20      , sd
=        3       )
sample_dataframe <-       as.data.frame(       samples )

ggplot  (       data =        sample_dataframe,
       aes     (       x =        samples )
       )       +
  geom_histogram(       aes     (       y =        ..density..)       ,
                 bins=       50      ,
                 color =       1       ,
                 fill =       'white' )       +
  geom_density(       )       +
  theme_minimal(       )
```



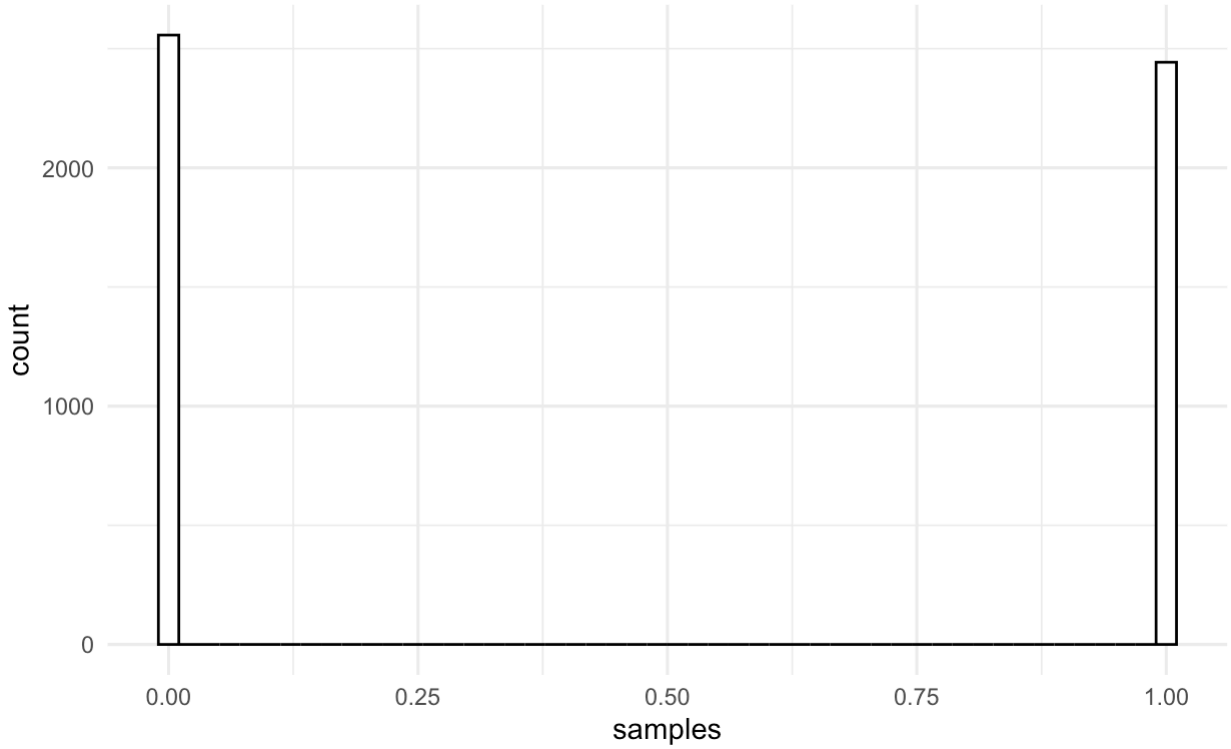
Binomial distribution

Not all samples can take any real value, as the normal distribution assumes. For instance, in an experiment there may be a binary condition: *pass/fail*; *right/wrong*; *left/right*; and so on. One intuitive way to think about such cases is as a coin flip (*heads*/*tails*). The binomial distribution is a discrete distribution of number of successes (e.g., *heads*) in a number of n independent experiments. It is defined by two parameters, the chances of success p and the number of experiments n. We will focus on the special case of a single experiment, i.e. $n = 1$ so we only need to keep track of p. This special case is called the Bernoulli distribution. The Bernoulli distribution has both an upper bound (number of trials) and a lower bound (0 successes).

Here's an illustration of 5000 samples drawn from a Bernoulli distribution with $p = 0.5$ (a fair coin, e.g.):

```
samples  <-       rbinom  (       n =        5000     , size =        1        , p
=        0.5     )        #5000 independent experiments with a single fair coin flip
sample_dataframe <-       as.data.frame(       samples )

ggplot  (       data =        sample_dataframe,
       aes     (       x =        samples )
       )       +
  geom_histogram(
                bins=       50       ,
                color =        1        ,
                fill =        'white' )        +
  theme_minimal(       )
```



Poisson distribution

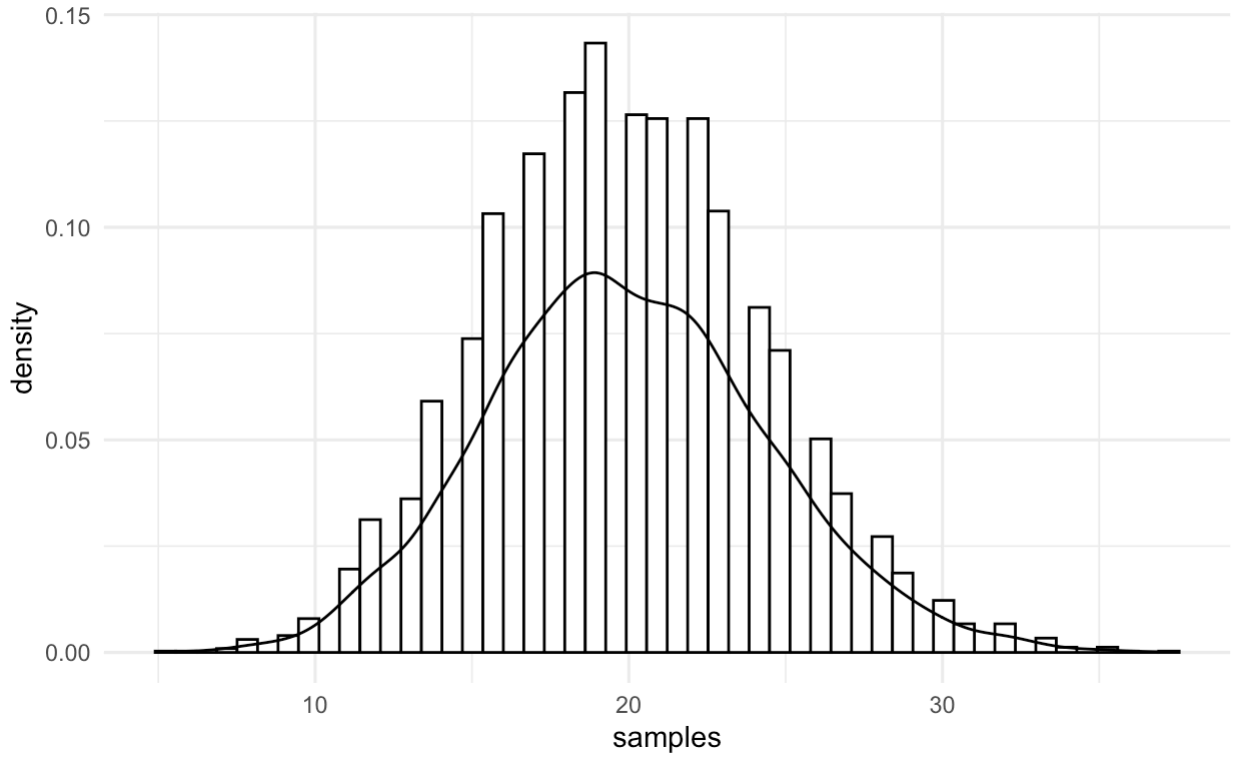
Not all samples can take any real value, as the Normal distribution assumes, nor be reduced to a (sequence) of discrete success/failure events, as binomially distributed events. Very often, we count discrete occurrences of events. If these events can be assumed to be well-approximated by a constant mean rate (usually denoted by λ), then the Poisson distribution may be worth considering. The Poisson distribution ranges from 0 to positive infinity. In other words, and by contrast to a binomial distribution, the Poisson distribution has no upper bound.

Here's an illustration of 5000 samples drawn from a Poisson distribution with $\lambda = 20$:

```
set.seed(       333     )
library (       ggplot2 )

samples   <-       rpois    (       n =         5000     , lambda =        20       )
sample_dataframe <-       as.data.frame(       samples )

ggplot   (       data =         sample_dataframe,
       aes      (       x =         samples )
       )        +
  geom_histogram(       aes      (       y =         ..density..)       ,
                 bins=       50       ,
                 color =        1       ,
                 fill =        'white' )        +
  geom_density(       )        +
  theme_minimal(       )
```



Corrections

If you spot a mistake or have suggestions, please get in touch with Thomas Brochhagen or create an issue

References