Lecture 5: Probablity distributions

Katie Schuler

2023-09-12

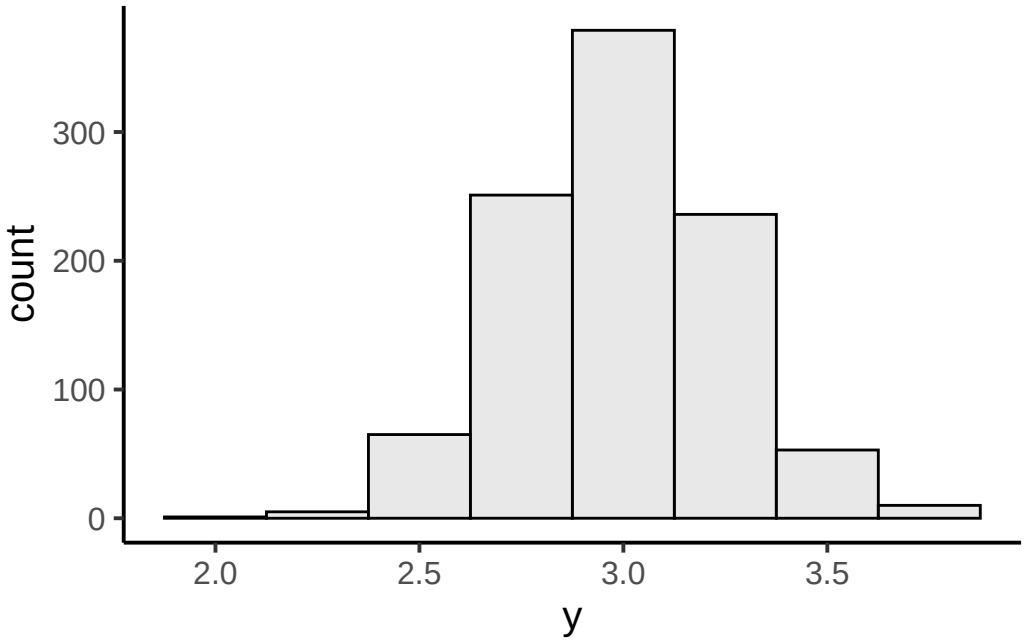
1 Exploring a simple dataset

```
# ------ setup for today's lecture notes ----- #

# suppress startup messages at package load
suppressPackageStartupMessages(library(tidyverse))

# set the theme for the plots
theme_set(theme_classic(base_size=15))

# generate 10000 y values with mean 3 and sd 0.25
data <- tibble(
    y=rnorm(1000, mean=3, sd=0.25)
)
```

We begin with the simplest possible dataset: suppose we measure a single quantity `y`. What can we do with these data?

We can create a visual summary of our dataset with a **histogram**. A histogram plots the distribution of a set of data, which allows us to get a quick visual of the data: formally we have plotted the the frequency distribution (count) of the data, but this also gives a sense of the central tendency and variability in our dataset.

```
ggplot(data=data, aes(x=y)) +
    geom_histogram(
        binwidth = 0.25,
        color="black", fill='lightgray', alpha=0.5
    )
```



2 Descriptive statistics

We can summarize (or describe) a set of data with **descriptive statistics**:

- Measures of **central tendency** describe where a central or typical value might fall (mean, median, mode)
- Measures of **variability** describe the dispersion or spread of values (variance, standard deviation, IQR)
- Measures of **frequency distribution** describe how frequently values occur (count)

R has built-in functions for descriptive statistics (we saw these in lecture 1):

```
data %>%
    summarise(
        n = n(),
        mean = mean(y),
        median = median(y),
        sd = sd(y),
        iqr_lower = quantile(y, 0.25),
        iqr_upper = quantile(y, 0.75)
    )
```

```
# A tibble: 1 x 6
      n  mean median    sd iqr_lower iqr_upper
  <int> <dbl>  <dbl> <dbl>     <dbl>     <dbl>
1  1000  2.99   2.99 0.253      2.82      3.16
```

Some stats are **paramteric** because they make assumptions about the the distribution of the data, and can therefore be computed from parameters:

- The mean and standard deviation assume the distribution is Gaussian and can therefore be computed via the following equations:

- **mean** μ
- **sd** σ

Other stats are **nonparametric** because they make minimal assumptions about the distribution of the data:

- The **median** is the 50th percentile, the value below which 50% of the data points fall.
- The **IQR** is the difference between the 25th and 75th percentiles (sometimes called the **50% coverage interval** because 50% of the data fall in this range).

- We can in principle calculate any arbitrary coverage interval. The 95% coverage interval — widely used in the sciences — is the difference between the 2.5 percentile and the 97.5 percentile, including all but 5% of the data.

3 Probability distributions

A **probability distribution** (aka probability density function) is a mathematical function that describes the probability of observing the different possible values of a variable (or variables). We will focus on univariate distributions in this class — probability distributions of just one random variable — but probability distributions can also be multivariate.

- One of the simplest probability distributions is the **uniform distribution**, where all possible values of a variable are equally likely. The probability density function for the uniform distribution is given by the following equation with two parameters (the boundaries, min and max):
  - $p(x) = \frac{1}{max - min}$
- One of the most useful probability distributions for our purposes is the **Gaussian (or Normal) distribution**. The probability density function for the Gaussian distribution is given by the following equation, with the parameters μ (mean) and σ (standard deviation):

- $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- The Gaussian distribution assumes that the distribution of a set of data takes a certain form (is unimodal, symmetric, etc).
- When values are sampled from a Gaussian distribution, 68% of the values will be within one standard deviation from the mean and 95% within two standard deviations from the mean.
- When computing the mean and standard deviation of a set of data, we are fitting a Gaussian distribution to the data.

4 Probability distributions with R

The probability distributions we've discussed so far are considered "parametric" because they are given by one or more parameters. When we use R's functions to generate values from these distributions, we provide these parameters as arguments. Base R has four functions we will use to generate values associated with a probability distribution. - `dnorm(mean=5, sd=1)` returns the height of the **probability density function** at the given values - `pnorm(5, mean=5, sd=1)` returns the **cumulative density function** (the probability that a random number from the distribution will be less than the given values) - `qnorm(0.8, mean=5, sd=1)` returns the value whose cumulative distribution matches the probability (**inverse of `p`**) - `rnorm(1000, mean=5, sd=1)` returns n **random numbers** generated from the distribution

To use another distribution, change the function's suffix to the name of the distribution and the parameters to those that define the distribution. For example, to generate n random numbers from a uniform distribution with a min of 1 and a max of 5, run `runif(n, min=0, max=1)`.

5 Nonparametric probability distributions

What if the data does not meet the assumptions of the Gaussian distribution? One option is to choose another parametric probability distribution (run `help(Distributions)` for a full list of available distributions). Another is to use a nonparametric approach, where the probability distribution is not determined by parameters but is instead determined by the data. - A **histogram** is actually a simple, nonparametric estimate of a probability distribution. To estimate the probability distribution that generated a set of data from a histogram, we modify the scale of the y-axis so that the total area of the bars is equal to 1. - **Kernel density estimation (KDE)** is another nonparametric method to estimate a probability distribution. KDE is like a smooth histogram, accomplished by placeing a kernel — a tiny Gaussian distribution — at each observed data point and summing across kernels. We can accomplish this in ggplot with the `geom_density()` geom.

Further reading and references

- Appendix A: Statistical Background in Modern Dive
- Ch 11: Modeling Randomness in Statistical Modeling

https://r4ds.hadley.nz/data-visualize#visualizing-distributions