Introduction to (Bayesian) Statistics

Motivation

Science and society are in crisis. The former battles with low statistical power (Cohen, 1962; Button et al., 2013) and failures to replicate (OSC, 2015) caused by a perverse "publish or perish" culture which encourages questionable research practices and overselling exploratory findings as the truth; society faces challenges on a grand scale such as climate change, the rise of nationalism and fascism, and increased loss of jobs due to automation. The "post-modern" society has been transformed into the "post-truth" society, with "leaders" such as Steve Bannon; experts are questioned, "alternative facts" proclaimed, and science waved aside.

But regardless of the detractors, rational thinking and the scientific method are the best tools available to make sense of the complex world we live in; and statistics is their steady companion; it's the Sancho Panza to the Don Quixote. Many scientists and students, however, have an insufficient grasp of statistics; but they are not to blame. Classical statistical concepts such as the p-value and confidence intervals are very unintuitive — even to seasoned statisticians.

In this workshop, I eschew the standard way of teaching statistics in psychology, i.e., introducing loosely connected tests in a cookbook-oriented fashion. Instead I provide an introduction to statistics from "first principles". All materials are available on https://github.com/fdabl/Intro-Stats. This handout provides a precis of the workshop and some practical exercises for you to work through in order to get a deeper understanding of the material.

History of Statistics

The history of statistics is intertwined with the history of probability. Probability was not put on solid mathematical foundation before 1933, when Andrey Kolmogorov. It first evolved through the applications of games of chance in the 18th century. Jakob Bernoulli ...

Mathematical concepts

Imagine having breakfast with a friend; clumsy as she is, her bread with butter slips out of her hands and falls on the floor. She is so embarrased that you, too, let your bread with butter slip out of your hands. What are the possible outcomes of this mishap?

Sets

The theory of probability is based on sets — unordered collections of things. Let $\mathcal{S}$ denote the set of all outcomes; call it the *sample space*. Either the bread falls on the floor with butter down (denote this by 1), or the bread falls on the floor butter up (denote this by 0). The elements or outcomes of our situation above are $\mathcal{S} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Let $\mathcal{A}$ denote the *event* that the bread lands butter down at least once; therefore, $\mathcal{A} = \{(0, 1), (1, 0), (1, 1)\}$. The complement of $\mathcal{A}$ is $\mathcal{A}^c = \{(0, 0)\}$ — the event that bread never lands butter down. Note that both $\mathcal{A}$ and $\mathcal{A}^c$ are subsets of $\mathcal{S}$, $\mathcal{A} \subset \mathcal{S}, \mathcal{A}^c \subset \mathcal{S}$. The intersection of $\mathcal{A}$ and $\mathcal{A}^c$ is $\mathcal{A} \cap \mathcal{A}^c = \emptyset$; the union is $\mathcal{A} \cup \mathcal{A}^c = \{(0, 0), (0, 1), (1, 0), (1, 1)\} = \mathcal{S}$.

Probability

Assuming that all outcomes are equally likely, we arrive at the *naive interpretation* of probability

$$P(\mathcal{A}) = \frac{|\mathcal{A}|}{|\mathcal{S}|}$$

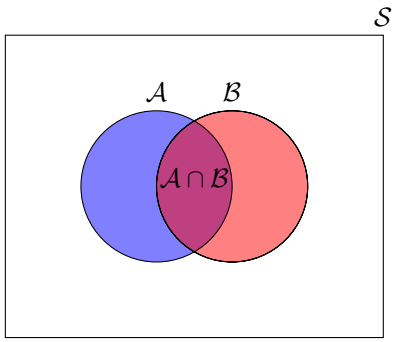
where $P(.)$ denotes the probability function. You can think of it as measuring the size of the set. In our example, $P(\mathcal{A}) = \frac{1}{2}$. More generally, we can introduce two axioms and call every $P(.)$ a probability function if it fulfills these two axioms.

$$P(\mathcal{S}) = 1,\ P(\emptyset) = 0$$
$$P(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} \mathcal{A}_i$$

From these two axioms, three important properties can be deduced.

$$P(\mathcal{A}^c) = 1 - P(\mathcal{A}^c)$$
$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$$
$$P(\mathcal{A}) \leq P(\mathcal{B}) \quad \text{if } \mathcal{A} \subset \mathcal{B}$$

Can you see how?



Random variables

Distributions

Statistical modeling

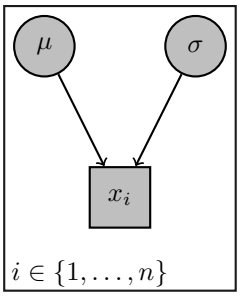
Assume I observe $x = \{x_1, x_2, \ldots, x_n\}$ data points. How would I best describe them to you? In order to reduce complexity, I introduce a statistical model which captures reality in a simplifying manner using a small set of *parameters*.

Normal model

The standard statistical model is the normal distribution. It's functional form is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{1}{2\sigma}(x-\mu)^2\big)$$

where μ describes the mean and σ the standard deviation of the data. Statistical models can be written as directedacyclical graphs, which provides us with powerful notation.

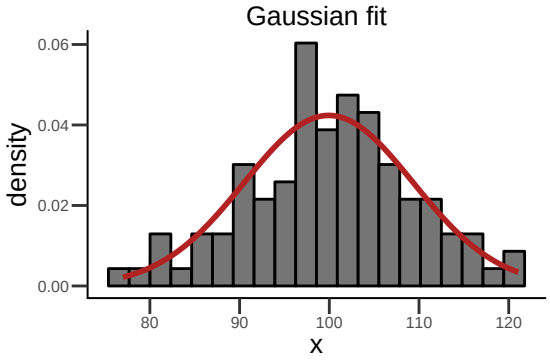


$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

$$x_i \sim \text{Norm}(\mu, \sigma)$$

We can fit such a model to the data.



In general, however, we cannot take for granted that the model is a good fit to the data. Model criticism and thinking hard about the data-generating process must be part of every statistical modeling enterprise.

Data analysis

You have many friends. People just want to hang out with you. However, this has a downside: lots of bread with butter falls on the floor! Last week alone you and your friends managed to have $n = 20$ slices of bread fall on the floor. The outcomes were

$$x = \{0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1\}$$

You ask yourself three questions: is a slice of bread with butter equally likely to fall on the floor butter up or butter down? What is the most likely value for the proportion? And what is the probability that the next slice of bread will fall butter down on the floor? The first question is one of **hypothesis testing**, the second one **parameter estimation**, and the third concerns itself with **model prediction**.

Before you can even start, you have to think about a statistical model; what process would best describe how the data came about? You start with a simpler problem; one slice of bread with butter falling down. You realize that the outcome is binary, $x \in \{0, 1\}$, and suggest the following model

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{otherwise} \end{cases}$$

But we have to deal with n slices of bread falling down, not just with one. You assume that the outcomes are (conditionally) independent; that is, assuming you know the true value of θ, you gain no information about

the next outcome by observing the last one. This allows you to use the multiplication rule of probability. You arrive at

$$\begin{aligned} f(x;\theta) &\overset{\text{i.i.d.}}{=} f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta) \\ &\vdots \\ &= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

You look at the last equation and you realize: the function only depends on the sum of the x_i! Quickly, you introduce a new random variable $Y = \sum_{i=1}^{n} x_i$ and rewrite the likelihood function as

$$f(y;\theta,n) = \theta^y (1-\theta)^{n-y}$$

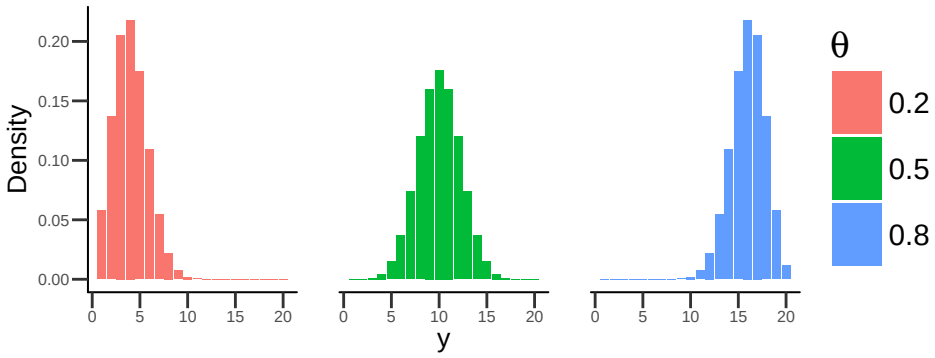
Thus instead of carrying the whole data vector x in your backpack, you can sufficiently summarize your data just with $n = 20$ and $y = 15$. Smirking, you make another observation. *I don't care about the order of the outcomes!*, you exclaim. It doesn't matter whether you observe

$$\begin{aligned} x &= \{0,1,1,1,1,1,1,1,0,1,1,0,0,0,1,1,1,1,1,1\} \\ x &= \{1,0,1,1,1,0,1,1,1,1,1,0,1,0,1,1,1,1,1,0\} \\ &\text{etc.} \end{aligned}$$

You realize that there are $n \cdot (n-1) \cdot \ldots \cdot (n-y+1)$ such sequences. Show that

$$n \cdot (n-1) \cdot \ldots (n-y+1) \overset{!}{=} \binom{n}{y}$$

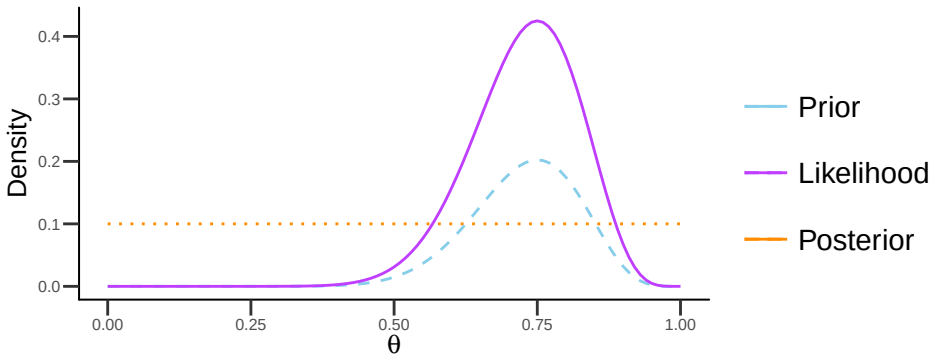
The likelihood function is a distribution over possible data patterns.



Parameter estimation proceeds by evoking Bayes' rule on the parameter θ

$$p(\theta|y,n) = \frac{f(y;\theta,n)p(\theta)}{\int_\Theta f(y;\theta,n)p(\theta)\mathrm{d}\theta}$$

We are agnostic towards any value of θ prior to data collection, and thus specify a uniform distribution.



A uniform distribution can be written as a Beta distribution with parameters $\alpha = 1, \beta = 1$.

$$p(\theta|\alpha, \beta) = \frac{1}{\mathcal{B}(a, b)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Now we just turn the Bayesian handle, plugging in the functional forms

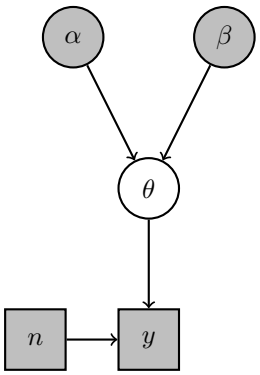
$$p(\theta|y, n) = \frac{f(y; \theta, n)p(\theta)}{\int_\Theta f(y; \theta, n)p(\theta)\mathrm{d}\theta}$$

$$p(\theta|y, n) = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{1}{\mathcal{B}(a,b)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_\Theta \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{1}{\mathcal{B}(a,b)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta}$$

With some rearranging we find that

$$p(\theta|y, n) \sim \text{Beta}(y + \alpha, n - y + \beta)$$

Can you write down the graphical model?



$\alpha = \beta = 1$

$n = 20$

$\theta \sim \text{Beta}(a, b)$

$y \sim \text{Binomial}(\theta, n)$

Hypothesis testing

Does the bread land equally often butter down or butter up? We compare two models that instantiate the hypotheses

$$M_0 : \theta = .5$$
$$M_1 : \theta \sim \text{Beta}(1, 1)$$

Bayesian hypothesis testing means evaluating the predictions of different models. The model which predicts the data most strongest gets the most boost in probability. We require a prior over θ for M_1 because otherwise the model would not make any predictions; $M_x : \theta \neq \theta$, as in classical statistics, is misspecified.

Applying Bayes' rule on models leads to

$$p(\theta|y, n, M_0) = \frac{f(y; \theta, n, M_0)p(\theta|M_0)}{\int_\Theta f(y; \theta, n, M_0)p(\theta|M_0)\mathrm{d}\theta}$$
$$p(\theta|y, n, M_1) = \frac{f(y; \theta, n, M_1)p(\theta|M_1)}{\int_\Theta f(y; \theta, n, M_1)p(\theta|M_1)\mathrm{d}\theta}$$