Models for proportions
Generative model

- The world is described by a model that governs the probabilities of observing different kinds of data.
Steps in Bayesian inference

- Specify a set of generative probabilistic models
- Assign a prior probability to each model
- Collect data
- Calculate the likelihood $P(\text{data}|\text{model})$ of each model
- Use Bayes’ rule to calculate the posterior probabilities $P(\text{model} | \text{data})$
- Draw inferences (e.g., predict the next observation)
**Likelihood P(d|Θ)**

- Data item $d$ is generated by a mechanism (model), parameters $Θ$ of which determine how probably different values of $d$ are generated, i.e., the distribution of $d$.

- An example:
  - Mechanism is drawing with replacement from a bucket of black and white balls, and the parameter $θ_b$ is the number of black balls, and the $θ_w$ is the number of white balls in a bucket:
    - $P(b|θ_b, θ_w) = θ_b / (θ_b + θ_w)$ and $P(w|θ_b, θ_w) = θ_w / (θ_b + θ_w)$.
  - In orthodox statistics, likelihood $P(D|Θ)$ is often seen as a function of $Θ$, a kind of $L_D(Θ)$. Whatever.
i.i.d.

- If the data generating mechanism depends on $\Theta$ only (and not on what has been generated before), the sequence of data data is called *independent and identically distributed*.

- Then

$$P (d_1, d_2, \ldots, d_n | \theta) = \prod_{i=1}^{n} P (d_i | \theta)$$

- And

  - order of $d_i$ does not matter.
  
  - $P (b, w, b, b, w | \theta) = P (b, b, w, w, w | \theta) = P (b | \theta) P (b | \theta) P (w | \theta) P (w | \theta) P (w | \theta)$
The Bernoulli model

- A model for i.i.d. binary outcomes (heads,tails), (1,0), (black, white), (true, false),...

- One parameter: $\Theta \in [0,1]$. For example: $P(d=\text{true} \mid \Theta) = \Theta$, $P(d=\text{false} \mid \Theta) = 1-\Theta$.
  - NB! The probabilities of $d$ being true are defined by the parameter $\Theta$. Parameters are not probabilities.
  - Black and white ball bucket as a Bernoulli model:
    - $\Theta$ is the proportion of black balls in a bucket $P(b \mid \Theta) = \Theta$.
    - $P(D \mid \Theta) = \Theta^{N_b} (1-\Theta)^{N_w}$, where $N_b$ and $N_w$ are numbers of black and white balls in the data $D$.
    - NB! $P(D \mid \Theta)$ depends on data $D$ through $N_b$ and $N_w$ only (=sufficient statistics)
Example

- You are installing WLAN-cards for different machines. You get the WLAN-cards from the same manufacturer, and some of them are faulty.
- We are asking the question: “Is the next WLAN-card we are installing going to work?”
- We are allowed to have background knowledge of these cards (they have been reliable/unreliable in the past, the manufacturing quality has gone up/down etc.)
Assessing models

- Let $A = \text{“The WLAN-card is not faulty”}$, and $B = \sim A$
- A proportion model can be understood as a bowl with labeled balls $(A, B)$
- each model $M(\Theta)$ is characterized by the number of $A$ balls, $\Theta$ is the proportion (Obs! Assume here that $\Theta$ is discrete, i.e., only consider $\Theta \in \{0, 0.1, 0.2, \ldots, 1\}$)
Our 11 models
Priors and the models

![Graph showing the distribution of M(\(\Theta\)) and P(M(\(\Theta\))) for different values of M(\(\Theta\)). The graph uses bars to represent the probability distribution for models A and B.]
The prior distribution $P(M(\Theta))$
Prediction by model averaging

- A Bayesian predicts by **model averaging**: the uncertainty about the model is taken into account by weighting the predictions of the different alternative models $M_i$ (=marginalization over the unknown)

$$P(X) = \sum_i P(X|M_i)P(M_i)$$
So: the predictive probability is...

- What is \( P(A) \), the probability that the next WLAN-card is not faulty?

\[
P(A) = P(A|M(0.0))P(M(0.0)) + P(A|M(0.1))P(M(0.1)) + \ldots + P(A|M(1.0))P(M(1.0))
\]

\[
= 0.0 + 0.02 + 0.03 + \ldots + 0.0 = 0.598
\]

- "Mean or average" model: \( \Theta = 0.598 \)
- 60/40 odds a priori
Enter some data ...

- Assume that I have installed three WLAN-cards: first was non-faulty (A), the two latter ones faulty (B), i.e., D={ABB}
- what are the updated (posterior) probabilities for the models M(Θ)?
- Enter Bayes, for example for M(0.6):

\[
P(M(0.6)|D) = \frac{P(D|M(0.6))P(M(0.6))}{P(D)}
\]

0.2
Calculating model likelihoods

- i.i.d.: we assume that the observations are independent given any particular model $M(\Theta)$
- $P(ABB \mid M(0.6)) = 0.6 \times 0.4 \times 0.4 = 0.096$
- This is repeated for each model $M(\Theta)$

To calculate the *likelihood* of a model, multiply the probabilities of the individual observations given the model.
Likelihood histogram $P(ABB|M(\Theta))$
Posterior = likelihood x prior

\[ P(\theta|D) \propto P(D|M(\theta))P(M(\theta)) \]
The normalizing factor $P(D)$

$$P(M(\theta)|D) = \frac{P(D|M(\theta))P(M(\theta))}{P(D)}$$

Calculate:

$$P(D|M(0.0))P(M(0.0)) = s_1$$
$$P(D|M(0.1))P(M(0.1)) = s_2$$
$$\ldots$$
$$P(D|M(1.0))P(M(1.0)) = s_{11}$$

Then:

$$P(D) = s_1 + s_2 + \ldots + s_{11}$$
Posterior distribution $P(M(\Theta)|D)$
Predictive probability with data D

- With data D, the prediction is based on averaging over the models $M(\Theta)$ weighted now by the posterior (instead of the prior used earlier) probability of the models:

$$P(X|D) = \sum_i P(X|M_i, D) P(M_i|D)$$
How did the probabilities change?

- The predictive probability $P(A \mid D) = P(A \mid ABB)$ that the next (fourth) WLAN-card is OK came down from the prior 60% to 52% (the change is not great because the data set is small).
Densities for proportions

- A richer set of models allows more precise proportion estimates, but comes with a cost: the amount of calculations necessary increase proportionally.
- We can move to consider infinite number of models:
  - Each model $\Theta$ is now a point on the interval from $[0,1]$.
  - We get a “smoothed” bar chart called a density $P(\Theta)$.
  - $\int P(\Theta) d\Theta = 1$
  - Only collections of models can have a probability $> 0$. 
Bayesian inference with densities?

- Using densities means that we no longer add probabilities, but calculate areas.
- To represent “infinite bar charts” we use curves that approximate the heights of bars.
- But how to predict with densities? We cannot go over all the individual models as we did in the discrete case.
- What about the prior?
Maximum likelihood

Given a data D, different values of \( \Theta \) yield different probabilities \( P(D|\Theta) \). The parameters that yield the largest probability of \( P(D|\Theta) \) are called maximum likelihood parameters for the data D.

- \( P(b,b,w,w,w|\Theta=0.7) = 0.7^2 0.3^3 = 0.1323 \)
- \( P(b,b,w,w,w|\Theta=0.1) = 0.1^2 0.9^3 = 0.00729 \)
- \( \arg\max_{\Theta} P(b,b,w,w,w|\Theta) = \arg\max_{\Theta} \Theta^2 (1-\Theta)^3 = ? \)
Likelihood \( P(b,b,w,w,w|\Theta) \)

- NB! Not a distribution, but a function of \( \Theta \).
ML-parameters for the Bernoulli model.  
(High school math refresher)

• So let us find ML-parameters for the Bernoulli model for the data with $N_b$ black balls and $N_w$ white ones.

$$P(D|\theta) = \theta^{N_b} (1-\theta)^{N_w},$$

so let us check when $P'(D|\theta) = 0, \theta \in ]0,1[.$

$$P'(D|\theta) = N_b \theta^{N_b-1} (1-\theta)^{N_w} + \theta^{N_b} N_w (1-\theta)^{N_w-1} - 1$$

$$= \theta^{N_b-1} (1-\theta)^{N_w-1} \left[ N_b (1-\theta) - \theta N_w \right]$$

$$= \theta^{N_b-1} (1-\theta)^{N_w-1} \left[ N_b - (N_b + N_w) \theta \right] = 0$$

$$\Leftrightarrow N_b - (N_b + N_w) \theta = 0 \Leftrightarrow \theta = \frac{N_b}{N_b + N_w}$$
But ML-parameters are too gullible

- Assume $D=(w,w)$, i.e., two white balls.
  - ML-parameter is $\Theta=0$.
  - Now $P(\text{next ball is black} \mid \Theta=0)= 0$.
  - Selecting ML parameters do not appear to be a rational choice.

- Be Bayesian:
  - Parameters are exactly the things you do not know for sure, so they have a (prior and posterior) distribution.
  - **Posterior distribution of the model is the goal of the Bayesian data-analysis.**
Predicting with posterior distribution

- Not a two phase process like in ML-case
  - first find ML parameters $\Theta$.
  - then use them to calculate $P(d|\Theta)$.

- Instead:
  \[
P(d|D) = \int_{\Theta} P(\theta, d|D)
  = \int_{\Theta} P(d|\theta, D)P(\theta|D)
  = \int_{\Theta} P(d|\theta)P(\theta|D)
\]

- Bayesian prediction uses predictions $P(d|\Theta)$ from all the models $\Theta$, and weighs them by the posterior probability $P(\Theta|D)$ of the models.
Posterior for Bernoulli parameter

- So likelihood \( P(D|\Theta) \) we can calculate.
- How about the prior \( P(\Theta) \)?
  - We should give a real number for each \( \Theta \).
    - One way out: as earlier, use a discrete set of parameters instead of continuous \( \Theta \). (Works, is flexible, but does not scale up well.)
    - Another way: Study calculus.
- And how about
  \[
P(D) = \int_{0}^{1} P(\theta) P(D|\theta) d\theta
  \]
Prior for Bernoulli model

- The form of the likelihood gives us a hint for a comfortable prior
  - \( P(D|\Theta) = \Theta^{Nb} (1-\Theta)^{Nw} \)
  - If we define the \( P(\Theta) = c \Theta^{\alpha-1} (1-\Theta)^{\beta-1} \),
    - \( c \) taking care that \( \int P(\Theta) d\Theta = 1 \), then
  - \( P(\Theta)P(D|\Theta) = c \Theta^{Nb+\alpha-1} (1-\Theta)^{Nw+\beta-1} \)

- Thus updating from prior to posterior is easy: just use the formula for the prior, and update exponents \( \alpha-1 \) and \( \beta-1 \) (conjugate prior).
$P(\Theta)$ of a form $c \Theta^{\alpha-1}(1-\Theta)^{\beta-1}$ is called Beta($\alpha, \beta$) distribution

- The expected value of $\Theta$ is $\alpha/(\alpha+\beta)$.
- The normalizing constant is

$$c = \frac{1}{\int_{0}^{1} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},$$

where $\Gamma$ is the gamma function, a continuous version of the factorial:

$$\Gamma(n) = (n-1)!$$
Posterior of the Bernoulli model

\[ P(\theta|D, \alpha, \beta) = \frac{\Gamma(\alpha + N_b + \beta + N_w)}{\Gamma(\alpha + N_b) \Gamma(\beta + N_w)} \theta^{\alpha + N_b - 1} (1 - \theta)^{\beta + N_w - 1} \]

- Thus, a posteriori, \( \Theta \) is distributed by Beta\((\alpha+N_b, \beta+N_w)\).

- And prediction:

\[
P(b|D, \alpha, \beta) = \int_0^1 P(b|\theta, D, \alpha, \beta) P(\theta|D, \alpha, \beta) \, d\theta \\
= \int_0^1 P(b|\theta) P(\theta|D, \alpha, \beta) \, d\theta = \int_0^1 \theta P(\theta|D, \alpha, \beta) \, d\theta \\
= E_P(\theta) = \frac{\alpha + N_b}{\alpha + N_b + \beta + N_w}.
\]
Bernoulli prediction

\[ P(b|D, \alpha, \beta) = \frac{\alpha + N_b}{\alpha + N_b + \beta + N_w} \]

- So \( P(b|w,w,\alpha=1,\beta=1) = (1+0) / (1+0+1+2) = 1/4 \).
  - Sounds more rational!
  - Notice how the *hyperparameters* \( \alpha \) and \( \beta \) act like extra counts.
  - That's why \( \alpha + \beta \) is often called “equivalent sample size”. The prior acts like seeing \( \alpha \) black balls and \( \beta \) white balls before seeing data.
Laplace smoothing = Beta(1,1)

- For Bayesian inference, we can use a single model $\Theta^*$ which is the mean of the Beta($\alpha,\beta$) density:
  - $\Theta^* = (\alpha + N_+)/(\alpha + N_+ + \beta + N_-)$

- E.g.: flip a coin 10 times, observe 7 heads (“success”). Assuming a uniform prior Beta(1,1), the posterior for the $\Theta$ becomes Beta(8,4), and hence the predictive probability of heads is $8/12 = 2/3$, or:
  - $\Theta^* = (7+1)/(10+2)$

- Also known as Laplace’s rule of succession or Laplace smoothing
Equivalent sample size

- Predictive probabilities change less radically when $\alpha + \beta$ is large
- Interpretation: before formulating the prior, one has experience of previous observations - thus with $\alpha + \beta$ one can indicate confidence measured in observations
- Called “prior sample size” or “equivalent sample size”
- Beta(1,1) is the uniform prior
- Beta(0.5,0.5) is the Jeffreys prior
One variable, more than two values

- Variable $X$ with possible values $1, 2, \ldots, n$.
- Parameter vector $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n)$ with $\Sigma \Theta_i = 1$.
- $P(X=i|\Theta) = \Theta_i$. Prior $P(\Theta) = \text{Dirichlet}(\Theta; \alpha_1, \alpha_2, \ldots, \alpha_n) = 
\frac{\Gamma \left( \sum_{i=1}^{n} \alpha_i \right)}{\prod_{i=1}^{n} \Gamma (\alpha_i)} \prod_{i=1}^{n} \theta_i^{\alpha_i-1}$
- Posterior $P(\Theta) = \text{Dir}(\Theta; \alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_n + N_n)$
- Prediction $P(x_i | D, \alpha) = \frac{\alpha_i + N_i}{\sum_{j=1}^{n} \alpha_j + N_j}$. 

One variable, more than two values

- Variable $X$ with possible values $1, 2, \ldots, n$.
- Parameter vector $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n)$ with $\Sigma \Theta_i = 1$.
- $P(X=i|\Theta) = \Theta_i$. Prior $P(\Theta) = \text{Dirichlet}(\Theta; \alpha_1, \alpha_2, \ldots, \alpha_n) = 
\frac{\Gamma \left( \sum_{i=1}^{n} \alpha_i \right)}{\prod_{i=1}^{n} \Gamma (\alpha_i)} \prod_{i=1}^{n} \theta_i^{\alpha_i-1}$
- Posterior $P(\Theta) = \text{Dir}(\Theta; \alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_n + N_n)$
- Prediction $P(x_i | D, \alpha) = \frac{\alpha_i + N_i}{\sum_{j=1}^{n} \alpha_j + N_j}$. 

One variable, more than two values

- Variable $X$ with possible values $1, 2, \ldots, n$.
- Parameter vector $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n)$ with $\Sigma \Theta_i = 1$.
- $P(X=i|\Theta) = \Theta_i$. Prior $P(\Theta) = \text{Dirichlet}(\Theta; \alpha_1, \alpha_2, \ldots, \alpha_n) = 
\frac{\Gamma \left( \sum_{i=1}^{n} \alpha_i \right)}{\prod_{i=1}^{n} \Gamma (\alpha_i)} \prod_{i=1}^{n} \theta_i^{\alpha_i-1}$
- Posterior $P(\Theta) = \text{Dir}(\Theta; \alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_n + N_n)$
- Prediction $P(x_i | D, \alpha) = \frac{\alpha_i + N_i}{\sum_{j=1}^{n} \alpha_j + N_j}$.