Fitting probability distributions to data

A: The normal distribution

Distributional modeling

A useful way to understand a data set:

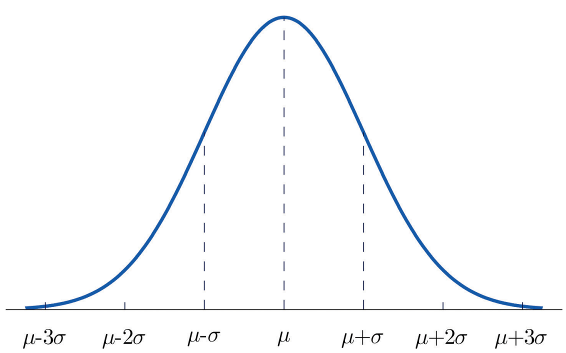
- Fit a probability distribution to it.
- Simple and compact.
- Captures the big picture while smoothing out the wrinkles in the data.
- In subsequent application, use distribution as a proxy for the data.

Which distributions to use?

> *There exist a few distributions of great universality which occur in a surprisingly large number of problems. The three principal distributions, with ramifications throughout probability theory, are the binomial distribution, the normal distribution, and the Poisson distribution.* – William Feller.

We'll see others as well. And for higher dimension, we'll use various combinations of 1-d models: **products** and **mixtures**.

The normal distribution

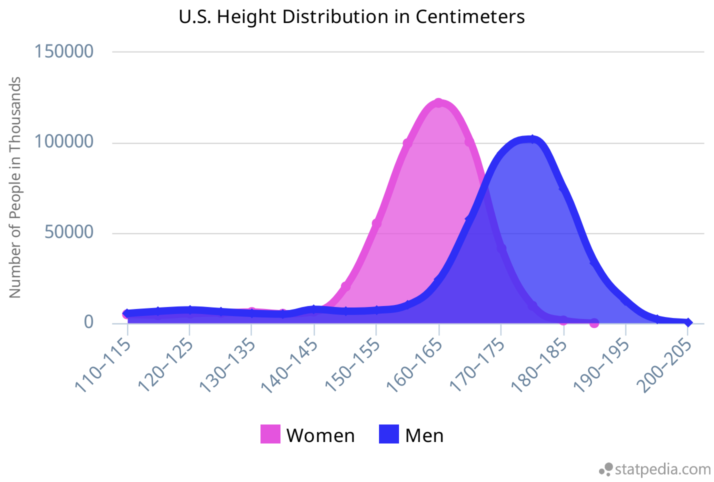


The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- 68.3% of the distribution lies within one standard deviation of the mean, $\mu \pm \sigma$
- 95.5% lies within $\mu \pm 2\sigma$
- 99.7% lies within $\mu \pm 3\sigma$

Gaussians are everywhere



Central Limit Theorem: Let $X_1, X_2, \ldots$ be independent with $\mathbb{E}X_i = \mu_i, \text{var}(X_i) = v_i$. Then

$$\frac{(X_1 + \cdots + X_n) - (\mu_1 + \cdots \mu_n)}{\sqrt{v_1 + \cdots + v_n}} \xrightarrow{d} N(0,1)$$

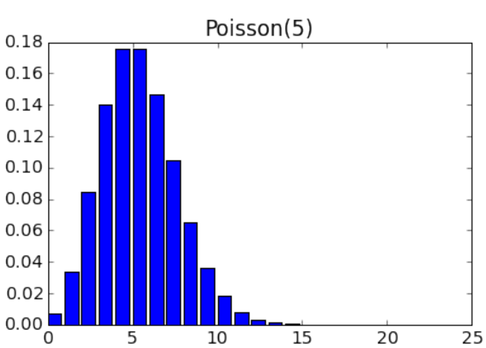
Fitting a Gaussian to data

Given: Data points $x_1, \ldots, x_n$ to which we want to fit a distribution.
What Gaussian distribution $N(\mu, \sigma^2)$ should we choose?

B: The Poisson distribution

The Poisson distribution

A distribution over the non-negative integers $\{0, 1, 2, \ldots\}$



Poisson(λ), with $\lambda > 0$:

$$\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- Mean: $\mathbb{E}X = \lambda$
- Variance: $\mathbb{E}(X - \lambda)^2 = \lambda$

How the Poisson arises

Count the number of events (collisions, phone calls, etc) that occur in a certain interval of time. Call this number X, and say it has expected value λ.

Now suppose we divide the interval into small pieces of equal length.

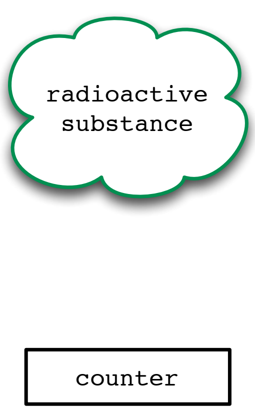
If the probability of an event occurring in a small interval is:

- independent of what happens in other small intervals, and
- the same across small intervals,

then $X \sim \text{Poisson}(\lambda)$.

Poisson: examples

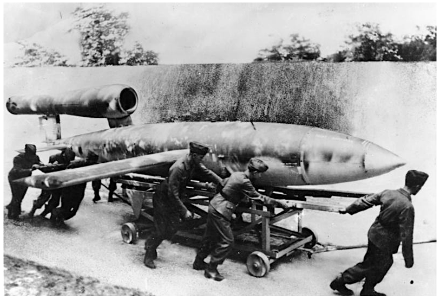
Rutherford's experiments with radioactive disintegration (1920)



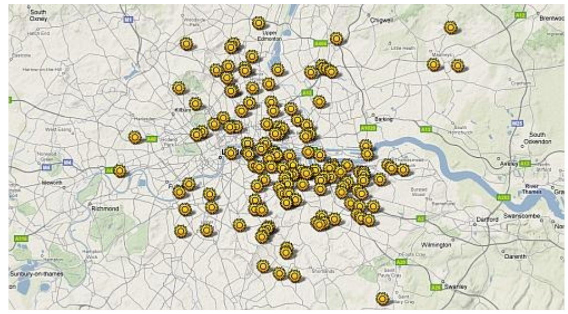
- $N = 2608$ intervals of 7.5 seconds
- $N_k = \#$ intervals with k particles
- Mean: 3.87 particles per interval

k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	57	203	383	525	532	408	273	139	45	43
P(3.87)	54.4	211	407	526	508	394	254	140	67.9	46.3

Flying bomb hits on London in WWII



Bundesarchiv, Bild 146-1975-117-26 / Lysiak / CC-BY-SA 3.0



- Area divided into 576 regions, each 0.25 km^2
- $N_k = \#$ regions with k hits
- Mean: 0.93 hits per region

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
P(0.93)	226.8	211.4	98.54	30.62	7.14	1.57

Fitting a Poisson distribution to data

Given samples $x_1, \ldots, x_n$, what Poisson(λ) model to choose?

C: Maximum likelihood estimation

Maximum likelihood estimation

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a class of probability distributions (Gaussians, Poissons, etc).

Maximum likelihood principle: pick the $\theta \in \Theta$ that makes the data maximally likely, that is, maximizes $\Pr(\textbf{data}|\theta) = P_\theta(\textbf{data})$.

Three steps:

1. Write down an expression for the **likelihood**, $\Pr(\text{data}|\theta)$.
2. Maximizing this is the same as maximizing its log, the **log-likelihood**.
3. Solve for the maximum-likelihood parameter θ.

Maximum likelihood estimation of the Poisson

$\mathcal{P} = \{\text{Poisson}(\lambda) : \lambda > 0\}$. We observe $x_1, \ldots, x_n$.

- Write down an expression for the **likelihood**, $\Pr(\text{data}|\lambda)$.

- Maximizing this is the same as maximizing its log, the **log-likelihood**.

- Solve for the maximum-likelihood parameter λ.

Maximum likelihood estimation of the normal

You see n data points $x_1, \ldots, x_n \in \mathbb{R}$, and want to fit a Gaussian $N(\mu, \sigma^2)$ to them.

- Maximum likelihood: pick μ, σ to maximize

$$\Pr(\text{data}|\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right)$$

- Work with the log, since it makes things easier:

$$\text{LL}(\mu, \sigma^2) = \frac{n}{2} \ln \frac{1}{2\pi\sigma^2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

- Setting the derivatives to zero, we get

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

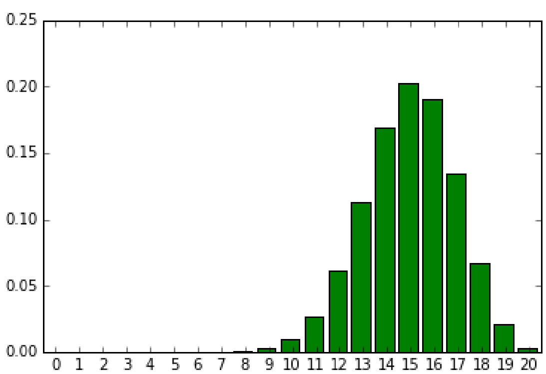
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

These are simply the empirical mean and variance.

D: The binomial distribution

The binomial distribution

Binomial(n, p): # of heads from n independent coin tosses of bias (heads prob) p.



For $X \sim \text{binomial}(n, p)$,

$$\mathbb{E}X =$$

$$\text{var}(X) =$$

$$\Pr(X = k) =$$

Fitting a binomial distribution to data

Example: Survey on food tastes.

- You choose 1000 people at random and ask them whether they like sushi.
- 600 say yes.

What is a good estimate for the fraction of people who like sushi? Clearly, 60%.

More generally, say you observe n tosses of a coin of unknown bias, and k come up heads. What distribution binomial(n, p) is the best fit to this data?

Maximum likelihood: a small caveat

You have two coins of unknown bias.

- You toss the first coin 10 times, and it comes out heads every time.
  You estimate its bias as $p_1 =$
- You toss the second coin 10 times, and it comes out heads once.
  You estimate its bias as $p_2 =$

Now you are told that one of the coins was tossed 20 times and 19 of them came out heads. Which coin do you think it is?

- Likelihood under p_1: $\Pr(19 \text{ heads out of } 20 \text{ tosses}|\text{bias} = 1) =$
- Likelihood under p_2: $\Pr(19 \text{ heads out of } 20 \text{ tosses}|\text{bias} = 0.1) =$

Laplace smoothing

A smoothed version of maximum-likelihood: when you toss a coin n times and observe k heads, estimate the bias as

$$p = \frac{k+1}{n+2}.$$

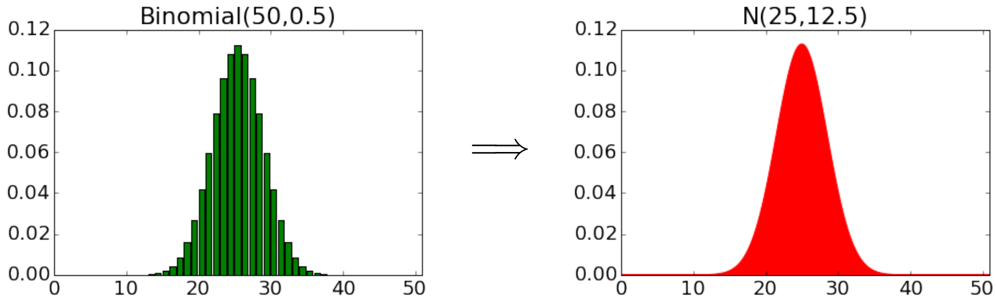
We will later justify this in a Bayesian setting.

Laplace's law of succession: What is the probability that the sun won't rise tomorrow?

- Let p be the probability that the sun won't rise on a randomly chosen day.
  We want to estimate p.
- For the past 5000 years ($= 1825000$ days), the sun has risen every day.
  Using Laplace smoothing, estimate

$$p = \frac{1}{1825002}.$$

Normal approximation to the binomial



When a coin of bias p is tossed n times, let S_n be the number of heads.

- We know S_n has mean np and variance $np(1-p)$.
- By central limit theorem: As n grows, the distribution of S_n looks increasingly like a Gaussian with this mean and variance, i.e.,

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1).$$

Poisson approximation to the binomial

Toss coins with bias $p_1, \ldots, p_n$ and let S_n be the number of heads.

Le Cam's inequality:

$$\sum_{k=0}^{\infty} \left| \Pr(S_n = k) - e^{-\lambda}\frac{\lambda^k}{k!} \right| \leq \sum_{i=1}^{n} p_i^2$$

where $\lambda = p_1 + \cdots + p_n$.

Poisson limit theorem: If all $p_i = \lambda/n$, then

$$S_n \xrightarrow{d} \text{Poisson}(\lambda).$$

Also called "the law of rare events".

E: The multinomial distribution

The multinomial distribution

Imagine a k-faced die, with probabilities $p_1, \ldots, p_k$.

Toss such a die n times, and count the number of times each of the k faces occurs:

$$X_j = \# \text{ of times face } j \text{ occurs}$$

The distribution of $X = (X_1, \ldots, X_k)$ is called the **multinomial**.

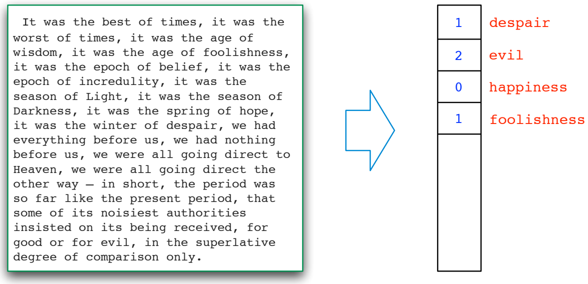
- Parameters: $p_1, \ldots, p_k \geq 0$, with $p_1 + \cdots + p_k = 1$.
- $\mathbb{E}X = (np_1, np_2, \ldots, np_k)$.
- $\Pr(n_1, \ldots, n_k) = \binom{n}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!},$$

the $\#$ of ways to place balls numbered $\{1, \ldots, n\}$ into bins numbered $\{1, \ldots, k\}$.

Example: text documents

Bag-of-words: vectorial representation of text documents.



- Fix V = some vocabulary.
- Treat words in document as independent draws from a multinomial over V:

$$p = (p_1, \ldots, p_{|V|}), \quad \text{such that} \quad p_i \geq 0 \text{ and } \sum_i p_i = 1$$

How would we estimate the parameters of a multinomial?

F: Alternatives to maximum likelihood?

Alternatives to maximum likelihood

Choosing a model in $\{P_\theta : \theta \in \Theta\}$ given observations $x_1, x_2, \ldots, x_n$.

- Maximum likelihood.
  The default, most common, choice.
- Method of moments.
  Pick the model whose moments $\mathbb{E}_{X \sim P_\theta} f(X)$ match empirical estimates.
- Bayesian estimation.
  Return the maximum a-posteriori distribution, or the overall posterior.
- Maximum entropy.
  We'll see this soon.
- Other optimization-based or game-theoretic criteria.
  As in generative adversarial nets, for instance.

Desiderata for probability estimators

Overall goal: Given data $x_1, \ldots, x_n$, want to choose a model P_θ, $\theta \in \Theta$.

- Let $T(x_1, \ldots, x_n)$ be some estimator of θ.
- Suppose $X_1, \ldots, X_n$ are i.i.d. draws from P_θ. Ideally $T(X_1, \ldots, X_n) \approx \theta$.

Some typical desiderata, if $X_1, \ldots, X_n \sim P_\theta$.

1. **Unbiased**: $\mathbb{E}\, T(X_1, \ldots, X_n) = \theta$.
2. **Asymptotically consistent**: $T(X_1, \ldots, X_n) \to \theta$ as $n \to \infty$.
3. **Low variance**: $\mathrm{var}(T(X_1, \ldots, X_n))$ is small.
4. **Computationally feasible**: Is $T(X_1, \ldots, X_n)$ easy to compute?

Do maximum-likelihood estimators possess these properties?

Are maximum likelihood estimators unbiased?

In general, no.

Example: Fit a normal distribution to observations $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

- Maximum likelihood estimate:

$$\widehat{\mu} = \frac{X_1 + \cdots + X_n}{n}$$
$$\widehat{\sigma}^2 = \frac{(X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2}{n}$$

- Can check that $\mathbb{E}[\widehat{\mu}] = \mu$ but

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{n-1}{n}\sigma^2.$$

Maximum likelihood: asymptotically consistent?

Not always, but under some conditions, yes.

Rough intuition:

- Given data $X_1, \ldots, X_n \sim P_{\theta^*}$, want to choose a model $P_\theta, \theta \in \Theta$.
- We pick the θ that maximizes

$$\begin{aligned}\frac{1}{n} LL(\theta) \;&=\; \frac{1}{n}\sum_{i=1}^{n} \ln P_\theta(X_i) \to \mathbb{E}_{X \sim P_{\theta^*}}[\ln P_\theta(X)] \\ &= \mathbb{E}_{X \sim P_{\theta^*}}[\ln P_{\theta^*}(X)] - K(P_{\theta^*}, P_\theta)\end{aligned}$$

Postscript: some other canonical distributions

We've seen the normal, Poisson, binomial, and multinomial.

Some others:

1. Gamma: two-parameter family of distributions over $\mathbb{R}^+$
2. Beta: two-parameter family of distributions over $[0, 1]$
3. Dirichlet: k-parameter family of distributions over the k-probability simplex

All of these are **exponential families** of distributions.