Fitting distributions to data

DSE 210

Distributional modeling

A useful way to understand a data set:

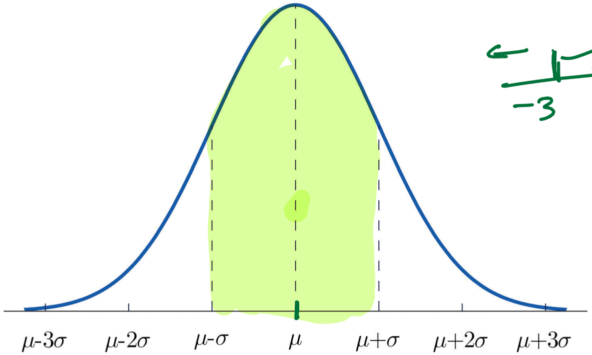
- Fit a probability distribution to it.
- Simple and compact.
- Captures the big picture while smoothing out the wrinkles in the data.
- In subsequent application, use distribution as a proxy for the data.

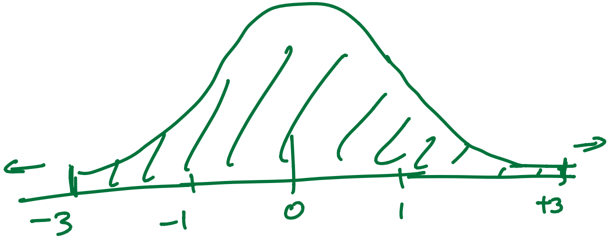
Which distributions to use?

> *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



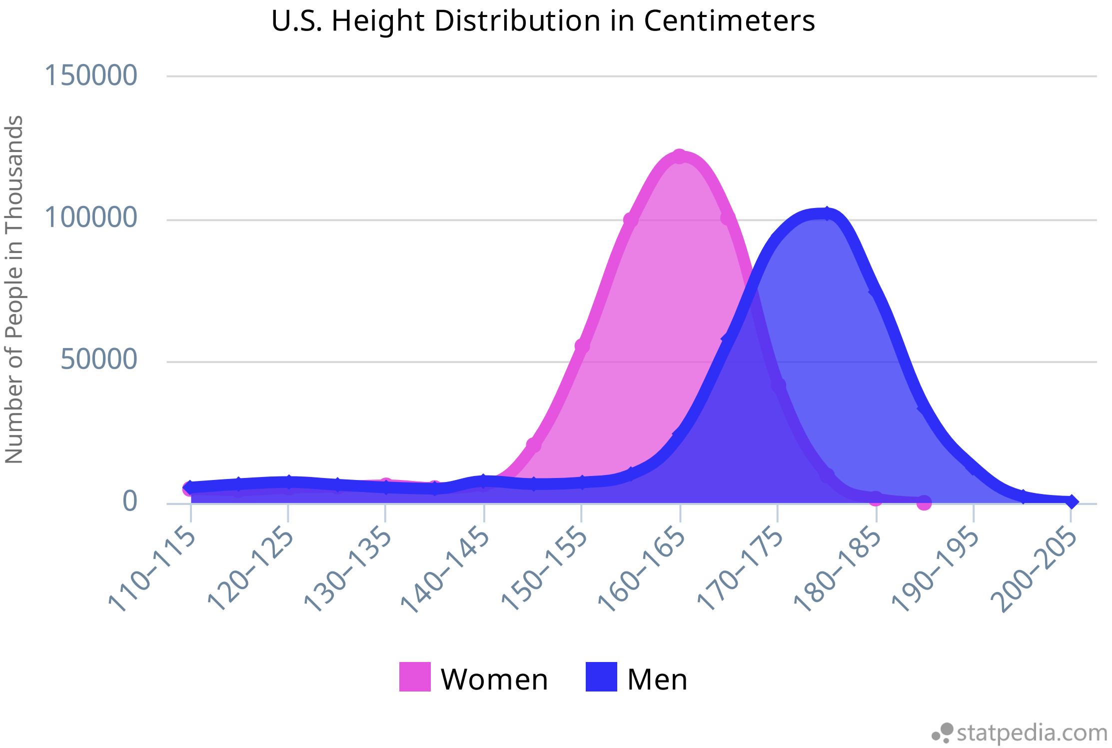


The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

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

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

Gaussians are everywhere



Fitting a Gaussian to data

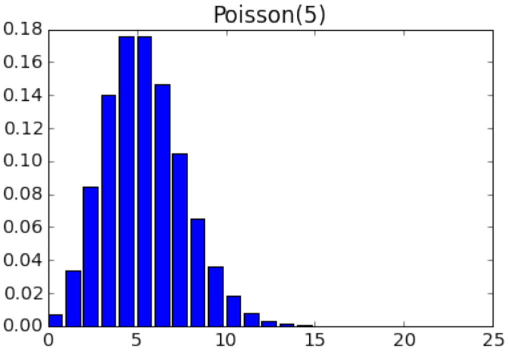
Given: Data points $x_1, \ldots, x_n$ to which we want to fit a distribution.
What Gaussian distribution $N(\mu, \sigma^2)$ should we choose?

$$\hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$$ "empirical mean"

$$\hat{\sigma}^2 = \frac{(x_1 - \hat{\mu})^2 + \cdots + (x_n - \hat{\mu})^2}{n}$$ "empirical variance"

The Poisson distribution

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



Poisson(λ), with $\lambda > 0$:

$$\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$ } sums to 1

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

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \cdot \Pr(X=k) = \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} = \cdots = \lambda$$

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:

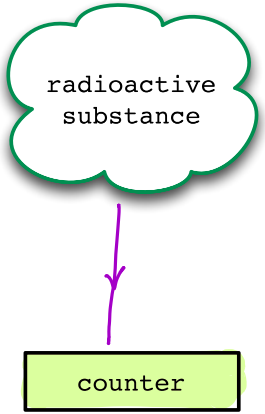
- independent of what happens in other small intervals, and
- the same across small intervals,

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

Poisson: examples

Rutherford's experiments with radioactive disintegration (1920)

$$3.87 = 0 \cdot \frac{57}{2608} + 1 \cdot \frac{203}{2608} + \cdots$$



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

there were 532 intervals (out of 2608) in which exactly 4 particles hit the counter

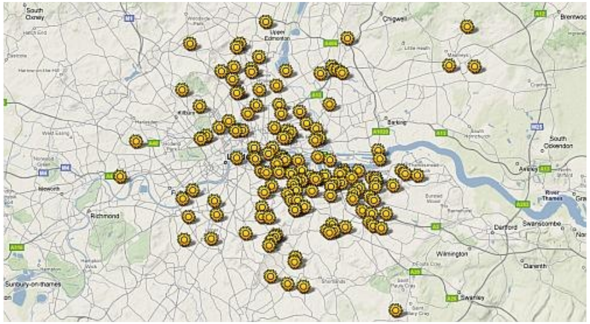
k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	57	203	383	525	532	408	273	139	45	43
$\underbrace{\text{P(3.87)}}_{\times 2608}$	54.4	211	407	526	508	394	254	140	67.9	46.3

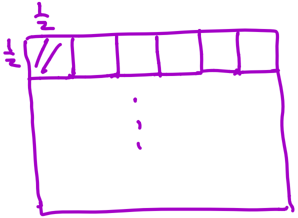
Flying bomb hits on London in WWII

Poisson with mean λ Poisson(λ)

V1 rockets

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





576 "boxes"

- 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

Poisson

Fitting a Poisson distribution to data

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

$$\lambda = \frac{x_1 + \dots + x_n}{n}$$

empirical mean

(since the Poisson has mean λ)

Is this really the best choice?

Why not use the empirical variance (since the Poisson also has variance λ)?

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 $\Pr(\textbf{data}|\theta) = P_\theta(\textbf{data})$.

Maximum likelihood estimation

$\ln(AB) = \ln A + \ln B$; $\ln e^x = x$; $\ln a^b = b \ln a$

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 $\Pr(\textbf{data}|\theta) = P_\theta(\textbf{data})$.

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

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

$$\Pr(\text{data}|\lambda) = \Pr(x_1 \ldots x_n|\lambda) = \underbrace{\prod_{i=1}^{n} \Pr(x_i|\lambda)}_{\text{product } \Pr(x_1|\lambda)\Pr(x_2|\lambda) \cdots \Pr(x_n|\lambda)} = \prod_{i=1}^{n}\left(e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right) = e^{-n\lambda}\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\,x_2!\cdots x_n!}$$

Prob of observing $x_1 \ldots x_n$ under the Poisson(λ) model

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

$$LL(\lambda) = \ln \Pr(\text{data}|\lambda) = -n\lambda + (x_1 + \cdots + x_n)\ln\lambda - \sum_{i=1}^{n}\ln(x_i!)$$

Pick the λ that maximizes this

- Solve for the maximum-likelihood parameter λ.

$$\frac{d}{d\lambda}LL(\lambda) = -n + \frac{x_1 + \cdots + x_n}{\lambda}$$

$\leftarrow$ set this to zero!

$$\lambda = \frac{x_1 + \cdots + x_n}{n}$$

this is the maximum-likelihood choice of λ given data $x_1, \ldots, x_n$

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

$\frac{d}{d\mu} LL$

$\frac{d}{d\sigma} LL$

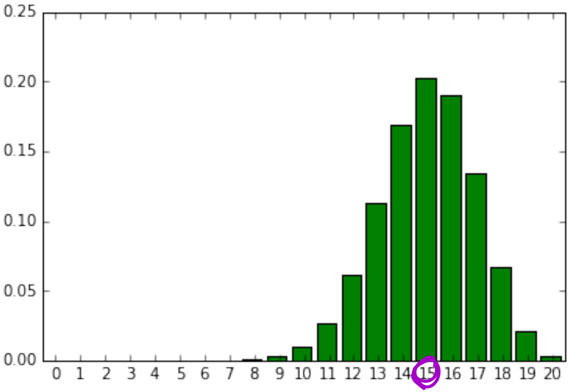
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

These are simply the empirical mean and variance.

The binomial distribution

Binomial(n, p): # of heads when n coins of bias (heads probability p) are tossed, independently.



$n = 20$

$p = 3/4$

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

$$\mathbb{E}X = np$$

$$\text{var}(X) = np(1-p)$$

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Fitting a binomial distribution to data

Example: Survey on food tastes.

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

What is a good estimate for the fraction of San Diegans 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? $p = k/n$

This is the max. likelihood choice.

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.

Max-likelihood estimate of bias: $p_1 = 1.0$ ← we need to smooth this estimate

- You toss the second coin 10 times, and it comes out heads once.

Max-likelihood estimate of bias: $p_2 = 0.1$

data

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?

Intuitively should be coin 1, which is strongly biased towards heads.

But:

$$\left.\begin{aligned} \Pr(\text{data} \mid p_1) &= p_1^{19}(1-p_1)^1 = 0 \\ \Pr(\text{data} \mid p_2) &= p_2^{19}(1-p_2)^1 > 0 \end{aligned}\right\}$$

suggests coin 2, which is ridiculous

Laplace smoothing

10 heads in a row

$p = 11/12$

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

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

Laplace smoothing

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

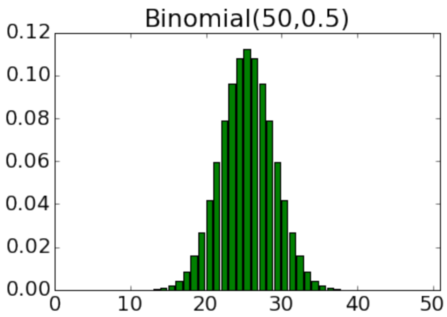
$$p = \frac{k+1}{n+2}. \qquad \frac{k+1/2}{n+1} \qquad \frac{k+2}{n+4}$$

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

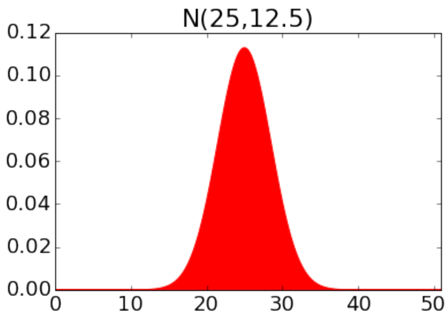
- Let p be the probability that the sun won't rise on a randomly chosen day. We want to estimate p.
- For the past 5000 years ($= 1825000$ days), the sun has risen every day. Using Laplace smoothing, estimate

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

Normal approximation to the binomial



$\Longrightarrow$



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

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

The multinomial distribution

Regular die: $k = 6$, $p_1 = p_2 = \cdots = p_k = 1/6$

Biased coin: $k = 2$, $p_1 = p$, $p_2 = 1-p$

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

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

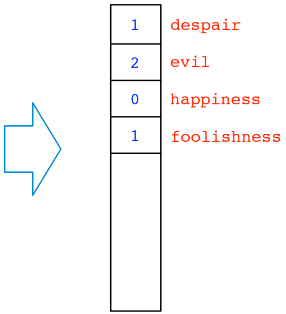
generalization of $\binom{n}{k} p^k (1-p)^{n-k}$

is the number 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.

```
It was the best of times, it was the
worst of times, it was the age of
wisdom, it was the age of foolishness,
it was the epoch of belief, it was the
epoch of incredulity, it was the
season of Light, it was the season of
Darkness, it was the spring of hope,
it was the winter of despair, we had
everything before us, we had nothing
before us, we were all going direct to
Heaven, we were all going direct the
other way – in short, the period was
so far like the present period, that
some of its noisiest authorities
insisted on its being received, for
good or for evil, in the superlative
degree of comparison only.
```



vector with $|V|$ entries, one per word in our vocabulary

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

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

Laplace smoothing

How would we estimate the parameters of a multinomial?

$$p_i = \frac{\#(\text{occurrences of word } i) + 1}{\text{total } \# \text{ words} + |V|}$$

Worksheet # 5

1, 2, 4, 5, 8