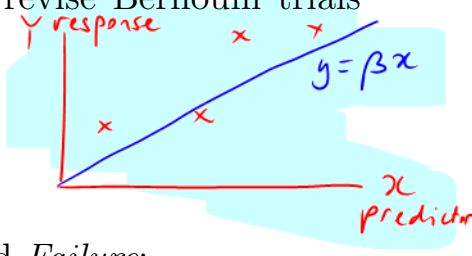
Chapter 3: Modelling with Discrete Probability Distributions

In Chapter 2 we introduced several fundamental ideas: hypothesis testing, likelihood, estimators, expectation, and variance. Each of these was illustrated by the Binomial distribution. We now introduce several other discrete distributions and discuss their properties and usage. First we revise Bernoulli trials and the Binomial distribution.

Regression models

$\mathbb{E}Y = \beta x$



Bernoulli Trials

A set of Bernoulli trials is a series of trials such that:

i) each trial has only 2 possible outcomes: *Success* and *Failure*;
ii) the probability of success, p, is constant for all trials;
iii) the trials are independent.

Examples: 1) Repeated tossing of a fair coin: each toss is a Bernoulli trial with $\mathbb{P}(\text{success}) = \mathbb{P}(\text{head}) = \frac{1}{2}$.

2) Having children: each child can be thought of as a Bernoulli trial with outcomes {girl, boy} and $\mathbb{P}(\text{girl}) = 0.5$.

3.1 Binomial distribution

Description: $X \sim \text{Binomial}(n, p)$ if X is the *number of successes out of a fixed number n of Bernoulli trials, each with $\mathbb{P}(\text{success}) = p$.*

Probability function: $f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ *for* $x = 0, 1, \ldots, n$.

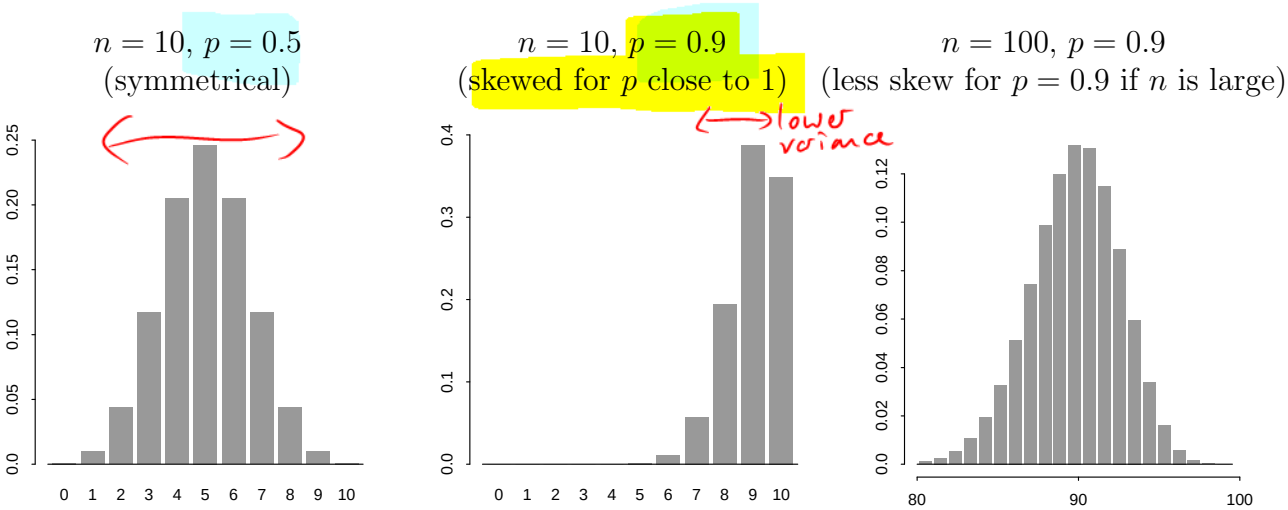
Mean: $\mathbb{E}(X) = np$.

Variance: $\text{Var}(X) = np(1-p)$.

Sum of independent Binomials: If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$, and if X and Y are *independent*, and if X and Y both share the same parameter p, then $X + Y \sim \textit{Binomial}(n + m,\ p)$.

$Var(X) = np(1-p)$ greatest when $p = 0.5$ & least when p is close to 0 or 1.

Shape: Usually symmetrical unless p *is close to 0 or 1.* Peaks at approximately np.



3.2 Geometric distribution

Like the Binomial distribution, the Geometric distribution is defined in terms of a sequence of Bernoulli trials.

- The Binomial distribution counts the *number of successes out of a fixed number of trials,* n.
- The Geometric distribution counts the *number of trials before the first success occurs.*

strictly before

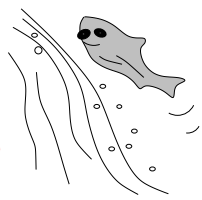
This means that the Geometric distribution counts the number of FAILURES before the first SUCCESS.

If every trial has probability p of success, we write: $X \sim \text{Geometric}(p)$.

Examples: 1) X =number of boys before the first girl in a family:

$X \sim \text{Geometric}(p = 0.5)$

2) Fish jumping up a waterfall. On every jump the fish has probability p of reaching the top.
Let X be the # failed jumps before the fish succeeds.



Then $X \sim \text{Geometric}(p)$.

Properties of the Geometric distribution

i) Description

$X \sim \text{Geometric}(p)$ if X is the *number of failures before the first success in a series of Bernoulli trials with* $\mathbb{P}(\textit{success}) = p$.

ii) Probability function

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

on formula sheet

no upper limit

$$\boxed{f_X(x) = \mathbb{P}(X = x) = (1-p)^x p \text{ for } x = 0, 1, 2, \ldots}$$

$x \geq 0$ wrong.
$x = 0, \ldots, n$ definitely wrong

Explanation:

$\mathbb{P}(X = x) = (1-p)^x * p$

need x failures... ... then one success, and you stop.

Difference between Geometric and Binomial: For the Geometric distribution, the trials must always occur in the order $\underbrace{FF\ldots F}_{x \text{ failures}}S$ single success at the end.

For the Binomial distribution, failures and successes can occur in any order: e.g. $FF\ldots FS$, $FSF\ldots F$, $SF\ldots F$, etc.

This is why the Geometric distribution has probability function

$$\mathbb{P}(x \text{ failures, } 1 \text{ success}) = (1-p)^x p,$$

no binomial coefficient because only one way of getting x fails and 1 success if S has to be at the end.

while the Binomial distribution has probability function

$$\mathbb{P}(x \text{ failures, } 1 \text{ success}) = \binom{\cancel{x+1}\ n}{x}(1-p)^x p.$$

out of n

iii) Mean and variance

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

$$\boxed{\begin{aligned} \mathbb{E}(X) &= \frac{1-p}{p} = \frac{q}{p} \\ \text{Var}(X) &= \frac{1-p}{p^2} = \frac{q}{p^2} \end{aligned}}$$

$\boxed{q = 1-p}$

On formula sheet.

Sum of Binomials: $X \sim Bin(n,p)$, $Y \sim Bin(m,p)$, indept

$X+Y \sim Bin(n+m, p)$

iv) Sum of independent Geometric random variables

If $X_1, \ldots, X_k$ are *independent*, and each $X_i \sim \text{Geometric}(p)$, then

$$X_1 + \ldots + X_k \sim \textit{Negative Binomial}(k, p). \qquad \textit{(see later)}$$

$X+Y$

↳ # fails before kth success, by defn.

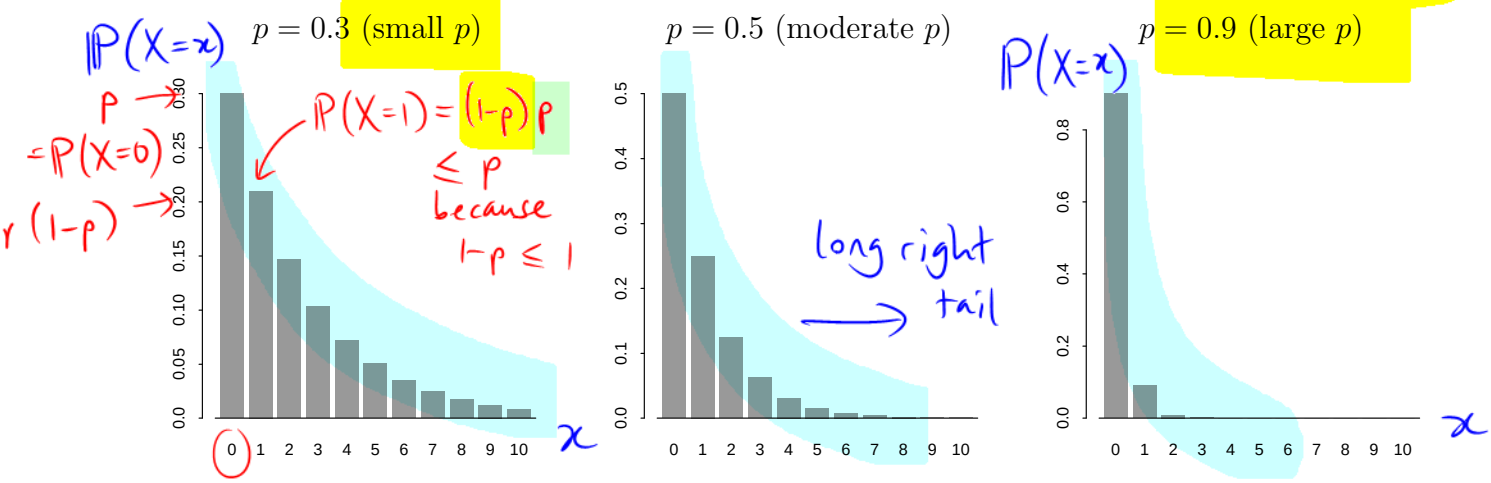
v) Shape

$\mathbb{P}(X=x) = (1-p)^x p \Rightarrow \mathbb{P}(X=0) = (1-p)^0 p = p$

Geometric probabilities are always greatest at $x=0$.
The distribution always has a long right tail (positive skew).

The length of the tail depends on p. For small p, there could be many failures before the first success, so the tail is long.

For large p, a success is likely to occur almost immediately, so the tail is short.



vi) Likelihood

For any random variable, the likelihood function is just the probability function expressed as a function of the unknown parameter. If:

- $X \sim \text{Geometric}(p)$;
- p is unknown;
- the observed value of X is x :

then the likelihood function is: $L(p\,;\,x) = p(1-p)^x$ for $0 \leq p \leq 1$.

Example: we observe a fish making 5 *failed* jumps before reaching the top of a waterfall. We wish to estimate the probability of success for each jump.

Then $L(p\,;\,5) = p(1-p)^5$ for $0 < p < 1$.

Maximise L with respect to p to find the MLE, $\hat{p}$.

MLEstimate $= \hat{p} = \frac{1}{5+1} = \frac{1}{6}$. MLEstimator $= \frac{1}{X+1}$.

For mathematicians: proof of Geometric mean and variance formulae (non-examinable)

We wish to prove that $\mathbb{E}(X) = \frac{1-p}{p}$ and $\mathrm{Var}(X) = \frac{1-p}{p^2}$ when $X \sim \text{Geometric}(p)$.

We use the following results:

$$\sum_{x=1}^{\infty} x q^{x-1} = \frac{1}{(1-q)^2} \qquad (\text{for } |q| < 1), \tag{3.1}$$

and

$$\sum_{x=2}^{\infty} x(x-1) q^{x-2} = \frac{2}{(1-q)^3} \qquad (\text{for } |q| < 1). \tag{3.2}$$

Proof of (3.1) and (3.2):

Consider the infinite sum of a geometric progression:

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} \qquad (\text{for } |q| < 1).$$

Differentiate both sides with respect to q:

$$\begin{aligned}
\frac{d}{dq}\left(\sum_{x=0}^{\infty} q^x\right) &= \frac{d}{dq}\left(\frac{1}{1-q}\right) \\
\sum_{x=0}^{\infty} \frac{d}{dq}(q^x) &= \frac{1}{(1-q)^2} \\
\sum_{x=1}^{\infty} x q^{x-1} &= \frac{1}{(1-q)^2}, \qquad \text{as stated in (3.1).}
\end{aligned}$$

Note that the lower limit of the summation becomes $x = 1$ because the term for $x = 0$ vanishes.

The proof of (3.2) is obtained similarly, by differentiating both sides of (3.1) with respect to q (Exercise).

Now we can find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^{\infty} x\mathbb{P}(X=x) \\
&= \sum_{x=0}^{\infty} xpq^x \qquad \text{(where } q=1-p\text{)} \\
&= p\sum_{x=1}^{\infty} xq^x \qquad \text{(lower limit becomes } x=1 \text{ because term in } x=0 \text{ is zero)} \\
&= pq\sum_{x=1}^{\infty} xq^{x-1} \\
&= pq\left(\frac{1}{(1-q)^2}\right) \qquad \text{(by equation (3.1))} \\
&= pq\left(\frac{1}{p^2}\right) \qquad \text{(because } 1-q=p\text{)} \\
&= \frac{q}{p}, \quad \text{as required.}
\end{aligned}$$

For $\mathrm{Var}(X)$, we use

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}\left\{X(X-1)\right\} + \mathbb{E}(X) - (\mathbb{E}X)^2 . \qquad (\star)$$

Now

$$\begin{aligned}
\mathbb{E}\{X(X-1)\} &= \sum_{x=0}^{\infty} x(x-1)\mathbb{P}(X=x) \\
&= \sum_{x=0}^{\infty} x(x-1)pq^x \qquad \text{(where } q=1-p\text{)} \\
&= pq^2\sum_{x=2}^{\infty} x(x-1)q^{x-2} \quad \text{(note that terms below } x=2 \text{ vanish)} \\
&= pq^2\left(\frac{2}{(1-q)^3}\right) \qquad \text{(by equation (3.2))} \\
&= \frac{2q^2}{p^2} .
\end{aligned}$$

Thus by $(\star)$,

$$\mathrm{Var}(X) = \frac{2q^2}{p^2} + \frac{q}{p} - \left(\frac{q}{p}\right)^2 = \frac{q(q+p)}{p^2} = \frac{q}{p^2},$$

as required, because $q+p=1$.

3.3 Negative Binomial distribution

The Negative Binomial distribution is a generalised form of the Geometric distribution:

- the Geometric distribution counts the number of *failures before the first success;*
- the Negative Binomial distribution counts the number of failures before the k'th success.

If every trial has probability p of success, we write: $X \sim NegBin(k, p)$.

Examples: 1) X =number of boys before the second girl in a family:

$X \sim NegBin(k=2, p=0.5)$.

2) Tom needs to pass 24 papers to complete his degree. He passes each paper with probability p, independently of all other papers. Let X be the #papers Tom fails in his degree.

Then $X \sim NegBin(k=24, p)$.

Properties of the Negative Binomial distribution

i) Description

$X \sim \text{NegBin}(k,\ p)$ if X is the number of failures before the k'th success in a series of Bernoulli trials with $\mathbb{P}(\text{success}) = p$.
$\Rightarrow$ indept

ii) Probability function

For $X \sim \text{NegBin}(k,\ p)$, Formula Sheet.

$$\boxed{f_X(x) = \mathbb{P}(X = x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \textit{for} \quad x = 0, 1, 2, \ldots}$$

Explanation:

- For $X = x$, we need x failures and k successes.
- The trials stop when we reach the k'th success, so the last trial must be a success.
- This leaves x failures and $k-1$ successes to occur in any order : a total of $x+k-1$ trials.

For example, if $x = 3$ failures and $k = 2$ successes, we could have:

$$FFFS\underline{S} \quad FFSF\underline{S} \quad FSFF\underline{S} \quad SFFF\underline{S}$$

So:

$$\mathbb{P}(X = x) = \underbrace{\binom{k+x-1}{x}}_{\text{\# ways of choosing } x \text{ trials to be failures out of } k+x-1 \text{ allocated trials in total.}} \times \overbrace{p^k}^{k \text{ successes}} \times \underbrace{(1-p)^x}_{x \text{ failures}}$$

iii) Mean and variance

For $X \sim \text{NegBin}(k,\, p)$,

$$\boxed{\begin{aligned} \mathbb{E}(X) &= \frac{k(1-p)}{p} = \frac{kq}{p} \\ \text{Var}(X) &= \frac{k(1-p)}{p^2} = \frac{kq}{p^2} \end{aligned}}$$

Formula sheet

$\text{Var}(X) = \frac{kq}{p} * \frac{1}{p}$
$= \mathbb{E}(X) * \frac{1}{p}$
$\Rightarrow \text{Var}(X) > \mathbb{E}(X)$

These results can be proved from the fact that the Negative Binomial distribution is obtained defined as the sum of k independent Geometric random variables:

$$\begin{aligned} X &= Y_1 + \ldots + Y_k, \quad \text{where each } Y_i \sim \text{Geometric}(p), \quad Y_i \text{ indept,} \\ \Rightarrow \quad \mathbb{E}(X) &= k\mathbb{E}(Y_i) = \frac{kq}{p}, \\ \text{Var}(X) &= k\text{Var}(Y_i) = \frac{kq}{p^2}. \end{aligned}$$

by independence.

iv) Sum of independent Negative Binomial random variables

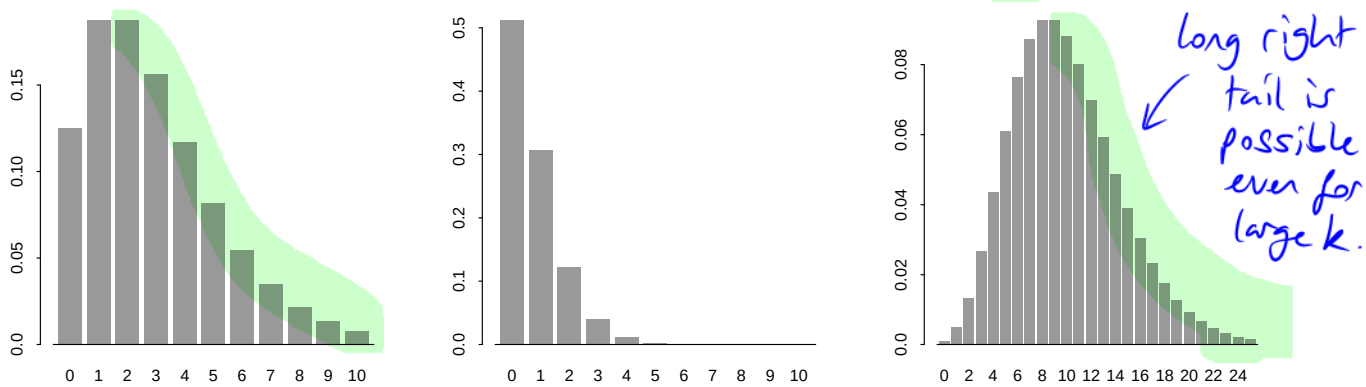
If X and Y are independent,
and $X \sim \text{NegBin}(k,\, p)$, $Y \sim \text{NegBin}(m,\, p)$, with the same value of p, then

$$X + Y \sim \text{NegBin}(k+m,\, p).$$

v) Shape (most exciting bit)

The Negative Binomial is flexible in shape. Below are the probability functions for various different values of k and p.

$k = 3,\ p = 0.5$ $\qquad$ $k = 3,\ p = 0.8$ $\qquad$ $k = 10,\ p = 0.5$



vi) Likelihood

As always, the likelihood function is the probability function expressed as a function of the unknown parameters. If:

- $X \sim \text{NegBin}(k,\ p)$;
- k is known;
- p is unknown;
- the observed value of X is x :

then the likelihood function is:

$$L(p\,;\,x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \text{for } 0 < p < 1.$$

Example: Tom fails a total of 4 papers before finishing his degree. What is his pass probability for each paper?

X = # failed papers before 24 passed papers.

Model $X \sim \text{NegBin}(24, p)$.

Observation: $X = 4$ failed papers.

Likelihood: $L(p\,;\,4) = \binom{24+4-1}{4} p^{24} (1-p)^4 = \binom{27}{4} p^{24} (1-p)^4$ for $0 < p < 1$.

Maximise L with respect to p to find the MLE, $\hat{p}$. $\quad \hat{p} = \dfrac{k}{X+k}$.

NonExaminable

3.4 Hypergeometric distribution: sampling without replacement

The hypergeometric distribution is used when we *are sampling without replacement from a finite population.*

i) Description

Suppose we have N objects:

- M of the N objects are *special;*
- the other $N - M$ objects are *not special.*

We remove n objects *at random without replacement.*

Let $X =$ *number of the* n *removed objects that are special.*

Then $X \sim$ *Hypergeometric*$(N,\ M,\ n)$.

Example: Ron has a box of Chocolate Frogs. There are 20 chocolate frogs in the box. Eight of them are dark chocolate, and twelve of them are white chocolate.

Ron grabs a random handful of 5 chocolate frogs and stuffs them into his mouth when he thinks that noone is looking. Let X be the number of dark chocolate frogs he picks.

Then $X \sim$ *Hypergeometric*$(N = 20,\ M = 8,\ n = 5)$.

ii) Probability function

For $X \sim \text{Hypergeometric}(N, M, n)$,

$$\boxed{f_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x = \max(0, n + M - N) \text{ to } x = \min(n, M).}$$

Explanation: We need to choose x special objects and $n-x$ other objects.

- Number of ways of selecting x special objects from the M available is: $\binom{M}{x}$.
- Number of ways of selecting $n-x$ other objects from the $N-M$ available is: $\binom{N-M}{n-x}$.
- Total number of ways of choosing x special objects and $(n-x)$ other objects is: $\binom{M}{x} \times \binom{N-M}{n-x}$.
- Overall number of ways of choosing n objects from N is: $\binom{N}{n}$.

Thus:

$$\mathbb{P}(X=x) = \frac{\textit{number of desired ways}}{\textit{total number of ways}} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

Note: We need $0 \le x \le M$ (number of special objects),
and $0 \le n-x \le N-M$ (number of other objects).
After some working, this gives us the stated constraint that

$$x = \max(0, n+M-N) \text{ to } x = \min(n, M).$$

Example: What is the probability that Ron selects 3 white and 2 dark chocolates?

X =# dark chocolates. There are $N = 20$ chocolates, including $M = 8$ dark chocolates. We need

$$\mathbb{P}(X=2) = \frac{\binom{8}{2}\binom{12}{3}}{\binom{20}{5}} = \frac{28 \times 220}{15504} = 0.397\,.$$

iii) Mean and variance

For $X \sim \text{Hypergeometric}(N, M, n)$,

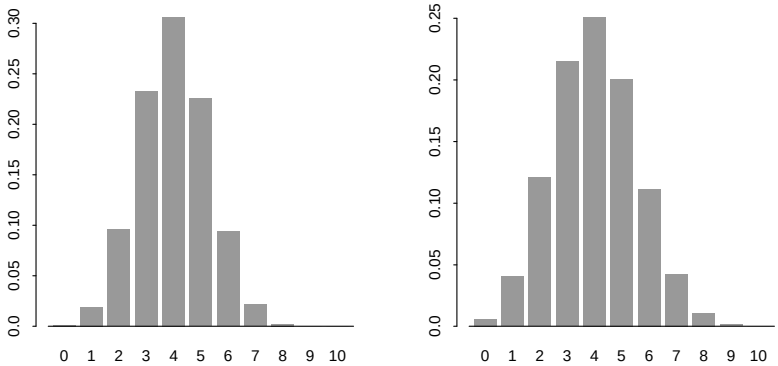
$$\boxed{\begin{aligned} &\mathbb{E}(X) = np \\ &\text{Var}(X) = np(1-p)\left(\tfrac{N-n}{N-1}\right) \end{aligned}} \qquad \text{where } p = \tfrac{M}{N}.$$

iv) Shape

The Hypergeometric distribution is similar to the Binomial distribution when n/N is small, because removing n objects does not change the overall composition of the population very much when n/N is small.

For $n/N < 0.1$ we often approximate the Hypergeometric(N, M, n) distribution by the ***Binomial***$(n, p = \frac{M}{N})$ ***distribution.***

Hypergeometric$(30, 12, 10)$ Binomial$(10, \frac{12}{30})$



Note: The Hypergeometric distribution can be used for opinion polls, because these involve sampling without replacement from a finite population.

The Binomial distribution is used when the population is sampled ***with replacement.***

As noted above, Hypergeometric$(N, M, n) \to$ Binomial$(n, \frac{M}{N})$ as $N \to \infty$.

A note about distribution names

Discrete distributions often get their names from mathematical power series.

- Binomial probabilities sum to 1 because of the Binomial Theorem:

$$\Big(p + (1-p)\Big)^n = \text{<sum of Binomial probabilities>} = 1.$$

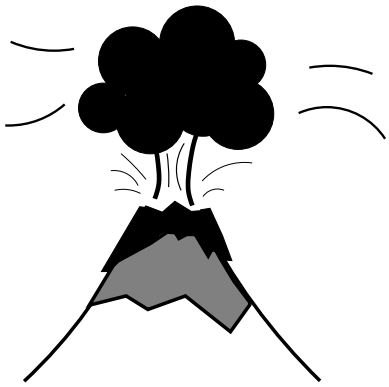
- Negative Binomial probabilities sum to 1 by the Negative Binomial expansion: i.e. the Binomial expansion with a negative power, $-k$:

$$p^k\Big(1 - (1-p)\Big)^{-k} = \text{<sum of NegBin probabilities>} = 1.$$

- Geometric probabilities sum to 1 because they form a Geometric series:

$$p\sum_{x=0}^{\infty}(1-p)^x = \text{<sum of Geometric probabilities>} = 1.$$

3.5 Poisson distribution



When is the next volcano due to erupt in Auckland?

Any moment now, unfortunately (give or take 1000 years or so).

A volcano could happen in Auckland this afternoon, or it might not happen for another 1000 years. Volcanoes are almost impossible to predict: they seem to happen completely at random.

discrete distn: values 0, 1, 2, ...

A distribution that counts the number of random events in a fixed space of time is the Poisson distribution.

How many cars will cross the Harbour Bridge today? $X \sim$ Poisson
How many road accidents will there be in NZ this year? $X \sim$ Poisson
How many volcanoes will erupt over the next 1000 years? $X \sim$ Poisson .

The Poisson distribution arose from a mathematical formulation called the Poisson Process, published by Siméon-Denis Poisson in 1837.

M. Poisson (Mr. Fish)

Poisson Process

(One example of how a Poisson distn is used, but the example of how it arises in nature).

The Poisson process counts the number of events occurring in a fixed time or space, when events occur independently and at a constant average rate.

Example: Let X be the number of road accidents in a year in New Zealand. Suppose that:

i) all accidents are independent of each other;

ii) accidents occur at a constant average rate of λ per year;

Greek letter "lambda"

iii) accidents cannot occur simultaneously.

Then the number of accidents in a year, X, has the exact distribution

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

ie. this is how the Poisson distn arises exactly in Nature.

Number of accidents in one year

Let X be the number of accidents to occur in one year:

The probability function for $X \sim \text{Poisson}(\lambda)$ is

$$\boxed{\mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \textit{for} \quad x = 0, 1, 2, \ldots}$$

LEARN

Average of λ/year
Average of $t\lambda$/tyrs

Number of accidents in t years

Let X_t be the number of accidents to occur in time t years.

Then $X_t \sim \text{Poisson}(\lambda t)$

and

$$\boxed{\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad \text{for} \quad x = 0, 1, 2, \ldots}$$

General definition of the Poisson process

Take any sequence of random events such that:

i) all events are *independent;*

ii) events occur at a *constant average rate of λ per unit time;*

iii) events *cannot occur simultaneously.*

Let X_t be the number of events to occur during time t.

Then $X_t \sim \text{Poisson}(\lambda t)$,

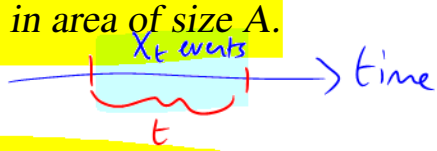
and

Poisson Process

unfolds over time

$$\boxed{\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad \text{for} \quad x = 0, 1, 2, \ldots}$$

Note: For a Poisson process in space, let X_A = *# events in area of size A.*
Then $X_A \sim$ *Poisson*(λA).



Example: X_A = number of raisins in a volume A of currant bun.

Where does the Poisson formula come from?

(Sketch idea, for mathematicians; non-examinable).
The formal definition of the Poisson process is as follows.

Definition: The random variables $\{X_t : t > 0\}$ form a Poisson process with rate λ if:

i) events occurring in any time interval are independent of those occurring in any other disjoint time interval;

ii)
$$\lim_{\delta t \downarrow 0} \left(\frac{\mathbb{P}(\text{exactly one event occurs in time interval}[t, t+\delta t])}{\delta t} \right) = \lambda\,;$$

iii)
$$\lim_{\delta t \downarrow 0} \left(\frac{\mathbb{P}(\text{more than one event occurs in time interval}[t, t+\delta t])}{\delta t} \right) = 0\,.$$

These conditions can be used to derive a partial differential equation on a function known as the *probability generating function* of X_t. The partial differential equation is solved to provide the form $\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$.

Poisson distribution ← more general use of the distn

now used in many other contexts

originally conceived through Poisson process

The Poisson distribution is not just used in the context of the Poisson process. It is also used in many other situations, often as a *subjective model* (see Section 3.7). Its properties are as follows.

i) Probability function

For $X \sim \text{Poisson}(\lambda)$,

LEARN

$$\boxed{f_X(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for} \quad x = 0, 1, 2, \ldots}$$

The parameter λ is called the rate of the Poisson distribution.

ii) Mean and variance

The mean and variance of the Poisson(λ) distribution are both λ :

$$\boxed{\mathbb{E}(X) = \lambda \quad \text{and} \quad Var(X) = \lambda \quad \text{when } X \sim \text{Poisson}(\lambda)}$$

LEARN

Notes:

1. It makes sense for $\mathbb{E}(X) = \lambda$: by definition, λ is the *average* number of events per unit time in the Poisson process.

2. The variance of the Poisson distribution increases with the mean (in fact, variance = mean). This is often the case in real life: there is more uncertainty associated with larger numbers than with smaller numbers.

(Remember NegBin : var > mean, even better for real life, often).

iii) Sum of independent Poisson random variables

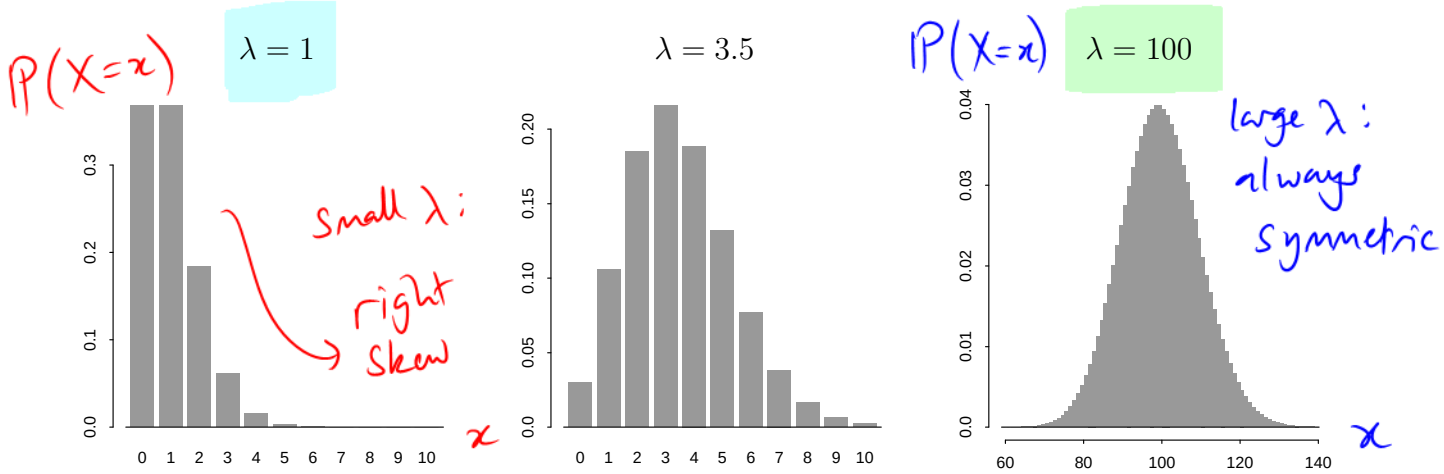
If X and Y are independent , and $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, then

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

iv) Shape

The shape of the Poisson distribution depends upon the value of λ. For small λ, the distribution has positive (right) skew. As λ increases, the distribution becomes more and more symmetrical, until for large λ it has the familiar bell-shaped appearance.

The probability functions for various λ are shown below.



v) Likelihood and Estimator Variance

As always, the likelihood function is the probability function expressed as a function of the unknown parameters. If:

- $X \sim \text{Poisson}(\lambda)$;
- λ is unknown;
- the observed value of X is $X = x$;

then the likelihood function is:

$$L(\lambda; x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } 0 < \lambda < \infty.$$

or $\lambda > 0$ is fine.

Example: 28 babies were born in Mt Roskill yesterday. Estimate λ.

Let X = # babies born in a day in Mt Roskill.

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

Obsn: $X = 28$ babies.

Likelihood: $L(\lambda; 28) = \frac{\lambda^{28}}{28!} e^{-\lambda}$ for $0 < \lambda < \infty$.

Maximise L with respect to λ to find the MLE, $\hat{\lambda}$.

We find that $\hat{\lambda} = x = 28$.

Similarly, the maximum likelihood *estimator* of λ is: $\hat{\lambda} = X$.

Thus the estimator variance is: $\text{Var}(\hat{\lambda}) = \text{Var}(X) = \lambda$ because $X \sim \text{Poisson}(\lambda)$ so $\text{Var}(X) = \lambda$.

Because we don't know λ, we have to *estimate* the variance:

$\hat{\text{Var}}(\hat{\lambda}) = \hat{\lambda} = 28$ here.

$\text{se}(\hat{\lambda}) = \sqrt{\hat{\text{Var}}(\hat{\lambda})} = \sqrt{28}$

approx 95% confidence interval for λ would be $\hat{\lambda} \pm 1.96\,\text{se}(\hat{\lambda})$

ie. $28 \pm 1.96\sqrt{28}$.

vi) R command for the p-value:

If $X \sim \text{Poisson}(\lambda)$, then the R command for $\mathbb{P}(X \leq x)$ is *ppois(x, lambda)*.

Proof of Poisson mean and variance formulae (non-examinable)

We wish to prove that $\mathbb{E}(X) = \text{Var}(X) = \lambda$ for $X \sim \text{Poisson}(\lambda)$.

For $X \sim \text{Poisson}(\lambda)$, the probability function is $f_X(x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$ for $x = 0, 1, 2, \ldots$

So

$$
\begin{aligned}
E(X) = \sum_{x=0}^{\infty} x f_X(x) &= \sum_{x=0}^{\infty} x \left(\frac{\lambda^x}{x!} e^{-\lambda} \right) \\
&= \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} e^{-\lambda} \qquad \text{(note that term for } x = 0 \text{ is 0)} \\
&= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} \quad \text{(writing everything in terms of } x-1\text{)} \\
&= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \qquad \text{(putting } y = x-1\text{)} \\
&= \lambda, \quad \text{because the sum=1 (sum of Poisson probabilities)}.
\end{aligned}
$$

So $\mathbb{E}(X) = \lambda$, as required.

For $\mathrm{Var}(X)$, we use:

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\
&= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}X)^2 \\
&= \mathbb{E}[X(X-1)] + \lambda - \lambda^2.
\end{aligned}
$$

$$
\begin{aligned}
\text{But } \mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x}{x!} e^{-\lambda} \\
&= \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} e^{-\lambda} \quad \text{(terms for } x = 0 \text{ and } x = 1 \text{ are 0)} \\
&= \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda} \quad \text{(writing everything in terms of } x-2\text{)} \\
&= \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \qquad \text{(putting } y = x-2\text{)} \\
&= \lambda^2.
\end{aligned}
$$

So

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X(X-1)] + \lambda - \lambda^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda, \quad \text{as required.}
\end{aligned}
$$

3.6 Likelihood and log-likelihood for n independent observations

So far, we have seen how to calculate the maximum likelihood estimator in the case of a

- $Y \sim \text{Binomial}(n, p)$ where n is known and p is to be estimated.
  Maximum likelihood estimator: $\widehat{p} = \frac{Y}{n}$.
- $Y \sim \text{Geometric}(p)$. *Maximum likelihood estimator:* $\widehat{p} = \frac{1}{Y+1}$.
- $Y \sim \text{NegBin}(k, p)$ where k is known and p is to be estimated.
  Maximum likelihood estimator: $\widehat{p} = \frac{k}{Y+k}$.
- $Y \sim \text{Poisson}(\lambda)$. *Maximum likelihood estimator:* $\widehat{\lambda} = Y$.

Question: What would we do if we had n independent observations, $Y_1, Y_2, \ldots, Y_n$?

Answer: As usual, the likelihood function is defined as the probability of the data, expressed as a function of the unknown parameter.
If the data consist of several independent observations, their probability is gained by multiplying the individual probabilities together.

Example: Suppose we have observations $Y_1, Y_2, \ldots, Y_n$ where each $Y_i \sim \text{Poisson}(\lambda)$, and $Y_1, \ldots, Y_n$ are independent. Find the maximum likelihood estimator of λ.

Before we start, what would you guess $\widehat{\lambda}$ to be in this situation?

Solution: For observations $Y_1 = y_1, \ldots, Y_n = y_n$, the likelihood is:

$$L(\lambda; y_1, \ldots, y_n) = \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) \text{ under parameter } \lambda$$

$$= \mathbb{P}(Y_1 = y_1 \cap Y_2 = y_2 \cap \cdots \cap Y_n = y_n)$$

$$= \mathbb{P}(Y_1 = y_1)\,\mathbb{P}(Y_2 = y_2) \cdots \mathbb{P}(Y_n = y_n) \text{ by independence}$$

$$\rightsquigarrow = \prod_{i=1}^{n} \left(\frac{\lambda^{y_i}}{y_i!} e^{-\lambda} \right) \quad \text{Poisson distn for each } Y_i$$

$$= \frac{1}{y_1!\, y_2! \cdots y_n!} (e^{-\lambda})^n \lambda^{y_1 + y_2 + \cdots + y_n}$$

$$\rightsquigarrow = K e^{-n\lambda} \lambda^{n\bar{y}} \quad \text{for } 0 < \lambda < \infty.$$

$n\bar{y} = y_1 + \dots + y_n$

So
$$L(\lambda\,; y_1, \dots, y_n) = K e^{-n\lambda}\,\lambda^{n\bar{y}},$$
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and $K = \prod_{i=1}^{n} \frac{1}{y_i!}$ is a constant that doesn't depend on λ.

Differentiate $L(\lambda\,; y_1, \dots, y_n)$ and set to 0 to find the MLE:

$$0 = \frac{d}{d\lambda} L(\lambda\,; y_1, \dots, y_n)$$
$$= K\left\{-n e^{-n\lambda}\lambda^{n\bar{y}} + (n\bar{y})\lambda^{n\bar{y}-1} e^{-n\lambda}\right\}$$ product differentiation
$$= K e^{-n\lambda}\lambda^{n\bar{y}-1}\left\{-n\lambda + n\bar{y}\right\}$$
$$\Rightarrow 0 = nK e^{-n\lambda}\lambda^{n\bar{y}-1}\{\bar{y} - \lambda\}.$$
$$\Rightarrow \lambda = \infty, \quad \lambda = 0, \quad \text{or} \quad \lambda = \bar{y}.$$

If we know that $L(\lambda\,; y_1, \dots, y_n)$ reaches a unique maximum in $0 < \lambda < \infty$, for example by reference to a graph, then we can deduce that the MLE is $\bar{y}$.

So the maximum likelihood estimator is:
$$\hat{\lambda} = \bar{Y} = \frac{Y_1 + \dots + Y_n}{n}.$$

Note: When $n = 1$, we get the same result as we had before: $\widehat{\lambda} = \frac{Y_1}{1} = Y_1$.

Log-likelihood

Instead of maximizing the likelihood function L to find the MLE, we often take logs and maximize the log-likelihood function, $\log L$. (***Note:*** $\log = \log_e = \ln$.) in Stats 210 always

There are several reasons for using the log-likelihood:

1. The logarithmic function $L \mapsto \log(L)$ is ***increasing,*** so the functions $L(\lambda)$ and $\log\{L(\lambda)\}$ will have the same maximum, $\hat{\lambda}$.
2. When there are observations $Y_1, \dots, Y_n$, the likelihood L is a product. Because $\log(ab) = \log(a) + \log(b)$, the log-likelihood converts the product into a sum. It is often easier to differentiate a sum than a product, so the log-likelihood is easier to maximize while still giving the same MLE.
3. If we need to use a computer to calculate and maximize the likelihood, there will often be numerical problems with computing the likelihood product, whereas the log-likelihood sum can be accurately calculated.
4. Because CLT applies to sums $\Rightarrow$ using $\log L$ will generate some super-fantastic Normal likelihood theory.

Example: Suppose we have observations $Y_1, Y_2, \ldots, Y_n$ where each $Y_i \sim \text{Poisson}(\lambda)$, and $Y_1, \ldots, Y_n$ are independent, as before. Use the ***log-likelihood function*** to find the maximum likelihood estimator of λ, and show that you get the same answer $\widehat{\lambda} = \overline{Y}$ as we obtained by maximizing the likelihood function directly.

Solution: *For observations $Y_1 = y_1, \ldots, Y_n = y_n$, the likelihood is:*

$$L(\lambda\,; y_1, \ldots, y_n) = \prod_{i=1}^{n} \left(\frac{\lambda^{y_i}}{y_i\,!} e^{-\lambda} \right) \quad \textit{(by independence)} \quad \text{(product)}$$

$$\textit{So } \log\{L(\lambda\,; y_1, \ldots, y_n)\} = \sum_{i=1}^{n} \log\left(\frac{\lambda^{y_i}}{y_i\,!} e^{-\lambda} \right) \qquad \text{sum}$$

$$= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{y_i\,!}\right) + \log\left(\lambda^{y_i}\right) + \log\left(e^{-\lambda}\right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{y_i\,!}\right) + y_i \log(\lambda) + (-\lambda) \right\}$$

$$= K' + \log(\lambda) \sum_{i=1}^{n} y_i - n\lambda \qquad \textit{where } K' \textit{ is a constant}$$

$$\log L(\lambda) \longrightarrow \quad = K' + \log(\lambda)\, n\overline{y} - n\lambda.$$

Differentiate w.r.t. λ *and set to 0 for the MLE:*

$$0 = \frac{d}{d\lambda} \log\{L(\lambda\,; y_1, \ldots, y_n)\}$$

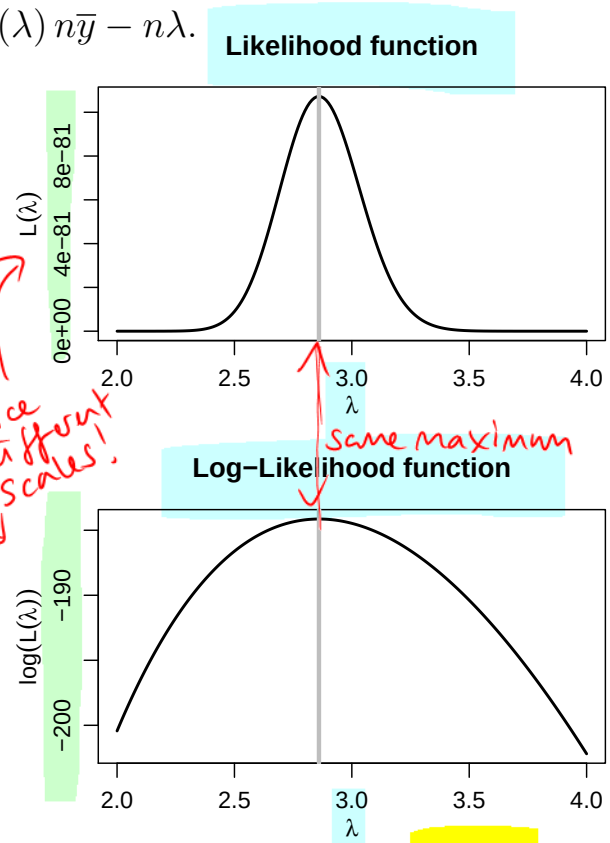
$$0 = \frac{d}{d\lambda} \{K' + \log(\lambda)\, n\overline{y} - n\lambda\}$$

$$\Rightarrow 0 = \frac{n\overline{y}}{\lambda} - n$$

$$\Rightarrow \widehat{\lambda} = \overline{y},$$

assuming a unique maximum in $0 < \lambda < \infty$.

So the MLE is $\widehat{\lambda} = \overline{Y}$ as before.



$L(\lambda)$ and $\log\{L(\lambda)\}$ for $n = 100$, $\overline{y} = 2.86$.

Read.

3.7 Subjective modelling

Most of the distributions we have talked about in this chapter are *exact* models for the situation described. For example, the Binomial distribution describes *exactly* the distribution of the number of successes in n Bernoulli trials.

However, there is often no exact model available. If so, we will use a ***subjective model.***

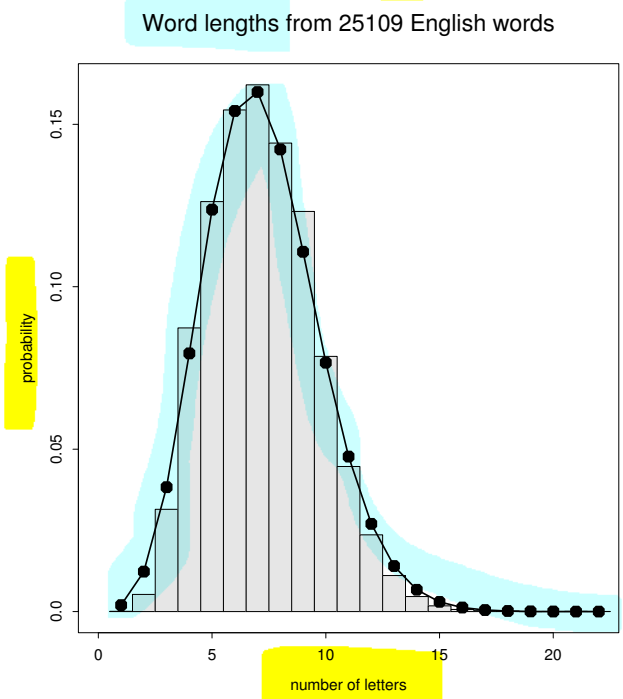
In a subjective model, we pick a probability distribution to describe a situation *just because it has properties that we think are appropriate to the situation, such as the right sort of symmetry or skew, or the right sort of relationship between variance and mean.*

Example: Distribution of word lengths for English words.

Let Y = *number of letters in an English word chosen at random from the dictionary.*

If we plot the frequencies on a barplot, we see that *the shape of the distribution is roughly Poisson.*

English word lengths: $Y - 1 \sim \text{Poisson}(6.22)$



The Poisson probabilities (with λ estimated by maximum likelihood) are plotted as points overlaying the barplot.

We need to use $Y \sim 1 + \text{Poisson}$ because Y cannot take the value 0.

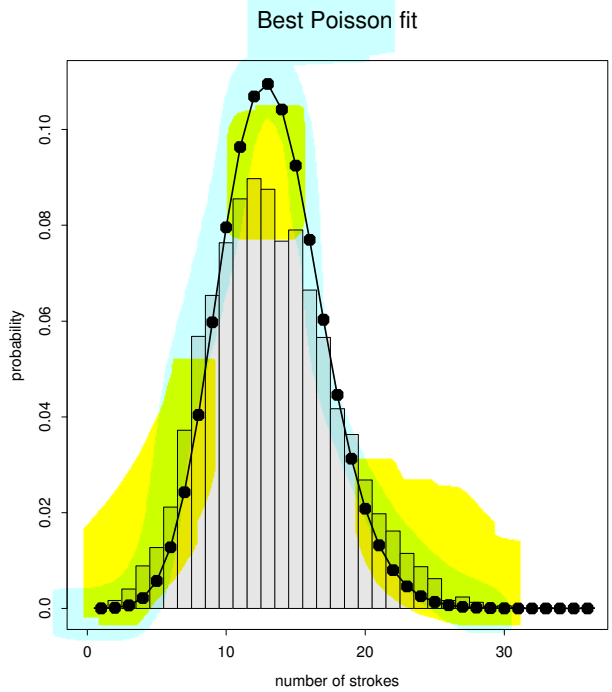
The fit of the Poisson distribution is *quite good.*

In this example we can not say that the Poisson distribution represents the number of events in a fixed time or space: *instead, it is being used as a subjective model for word length.*

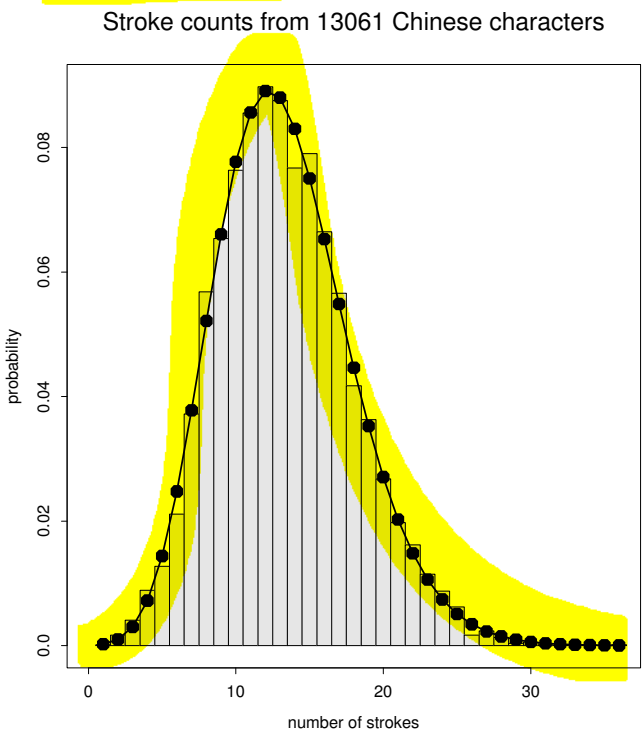
Can a Poisson distribution fit any data? The answer is *no: in fact the Poisson distribution is very inflexible.*

Here are stroke counts from 13061 Chinese characters. Y is the number of strokes in a randomly chosen character. The best-fitting Poisson distribution *(found by MLE)* is overlaid.

The fit of the Poisson distribution is *awful.*



It turns out, however, that the Chinese stroke distribution is well-described by *a Negative Binomial model.*



The best-fitting Negative Binomial distribution *(found by MLE)* is $\text{NegBin}(k = 23.7, p = 0.64)$. The fit is *very good.*

However, Y does not represent the number of failures before the k'th success: the Negative Binomial is a *subjective model.*

x : predictor
Y : response
} want to get an easily-understood expression for how Y depends on x.

3.8 Statistical regression modelling

NOT change of variable:
NOT $Y = g(X)$ e.g. $Y = X^2$ } e.g. mr chance

Statistical regression modelling is a fundamental technique used in data analysis in science and business. In this section we give an introduction to the idea of regression modelling, using the simplest example of modelling a straight line through the origin of a scatterplot.

In statistical regression, we explore the relationship between two variables.

One variable, x, is typically under our control.

We select several different values of x. At each value of x, we make measurements of the other variable, Y.

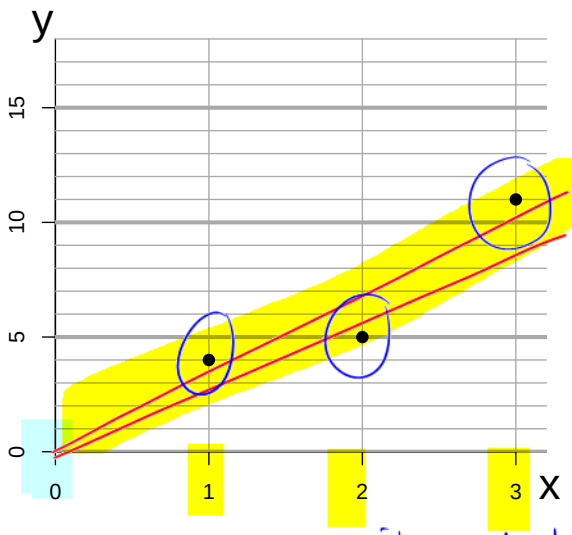
The other variable, Y, is regarded as random.

The distribution of Y ***depends upon the value of*** x at which we measure it.

We write (x_i, Y_i) for the i'th pair of measurements, where $i = 1, 2, \ldots, n$.

After the measurements are observed, we use lower-case letters and write (x_i, y_i).

Example: Where would you draw the best-fit line through the origin?



} different possible values of the slope, β.

- x is called the ***predictor variable,*** because it predicts the distribution of Y.
- Y is called the ***response variable,*** because it is observed in response to selecting a particular value of x.
- You might sometimes see x called the 'independent variable' and Y called the 'dependent variable'. Although it is widely used, this terminology is confusing because x is not independent of Y in a statistical sense. Most statisticians avoid this language and use the terms ***predictor*** and ***response*** instead.

How does the distribution of Y depend upon x?

In regression modelling, we generally assume that MEAN of Y's distribution, $\mathbb{E}Y$, has some relationship with the value of x.

The simplest regression model is a straight line through the origin. In this model, we assume that:

$$\mathbb{E}(Y) = \beta x,$$

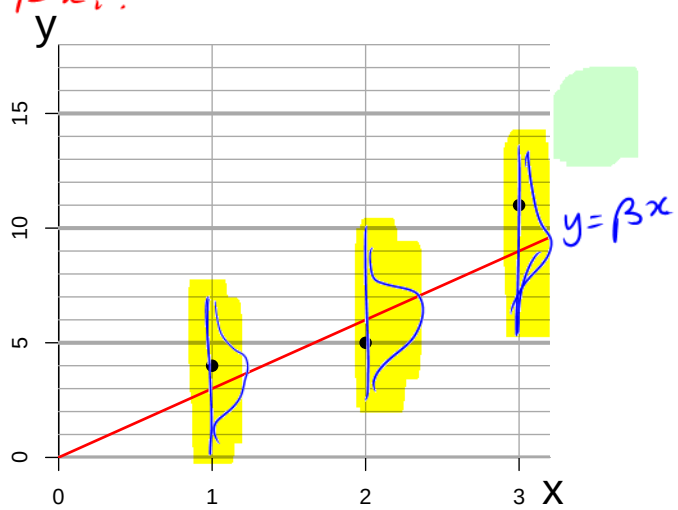
where the ***slope parameter*** β is what we want to estimate.

More specifically, in each of the pairs (x_i, Y_i) for $i = 1, \ldots, n$, we assume the same relationship $\mathbb{E}(Y_i) = \beta x_i$. ie. β doesn't change.

- The parameter β stays the same for all $i = 1, \ldots, n$. It gives the slope of the best-fit line through the origin.
- The mean of Y changes as x changes. When x is large, Y has a larger mean than when x is small (assuming $\beta > 0$). The mean of Y is $\mathbb{E}(Y_i) = \beta x_i$.

The line has equation $y = \beta x$. (here, $\beta = 3$).

The line shows the MEAN of the distn of Y at each point x.



Why do we want to fit a line to these points?

Our main interest is in the ***relationship*** between x and Y. In regression through the origin, this relationship is captured by the slope of the line, β.

Example:

- x represents some level of experience: e.g. $x=1$ for children in year 1 of school, $x=2$ for 2nd-years, etc.
- Y represents some sort of achievement: e.g. Y could be a reading score or numeracy score.
- The slope of the line, β, tells us about the improvement in children's scores from one year to the next.
- The school needs to prove that β is sufficiently high, and not 0 or negative!

Statistical model for Y

So far we have only specified the relationship between x and the mean of Y:

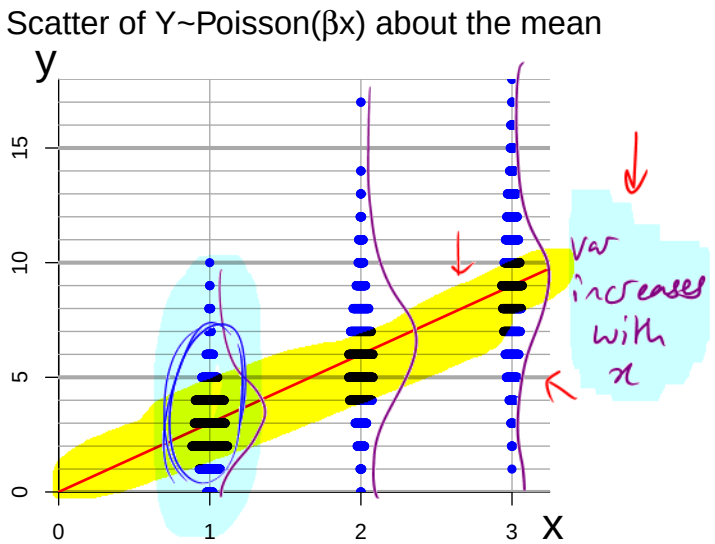
$$\mathbb{E}(Y_i) = \beta x_i.$$

In order to estimate the slope β, we need to specify a model for the whole distribution of Y.

Example 1: Let $Y_i \sim \text{Poisson}(\beta x_i)$. Then $\mathbb{E}(Y_i) = \beta x_i$ for all $i = 1, \ldots, n$.
This model could be suitable if Y_i measures a ***count*** of some item that depends upon x_i and has no upper limit. In the school example, Y_i could be a number of achievements or credits accumulated over the years.

As another example, a university student could create a model in which x_i is the percentage course credit awarded for an assignment, and Y_i is the time in hours spent on the assignment. We would expect more time to be spent on assignments with higher credit, but there will be randomness (scatter) about the straight-line relationship. A student might use this model to *look for outliers, to decide whether a particular assignment takes an unreasonably long time for the amount of credit awarded!*

Properties of $Y_i \sim \text{Poisson}(\beta x_i)$:



- Y_i takes values 0, 1, ... with no upper limit.
- $\text{Var}(Y_i) = \mathbb{E}(Y_i) = \beta x_i$, so variance increases with mean.

It is often appropriate to allow the variance to increase with the mean. If you estimate an assignment is going to take you 1 hour, you are unlikely to be wrong by 10 hours: there is ***low variance*** about a mean of $\mathbb{E}(Y) = 1$. Conversely, if you estimate the assignment will take 100 hours, it could easily take 10 hours more or less: there is ***higher variance*** about the mean of $\mathbb{E}(Y) = 100$.

Although the Poisson distribution allows variance to increase with the mean, it also makes a very specific assumption about the increase: under $Y \sim$ Poisson model, the variance of Y EQUALS the mean.

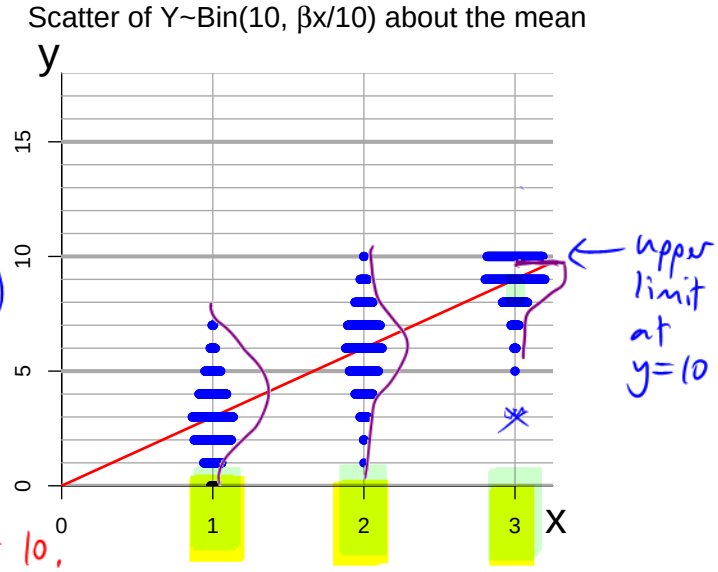
This assumption is often good enough, but equally it is often too restrictive. The usual problem is that the Poisson distribution doesn't allow the variance to increase enough as the mean gets larger. If so, modellers often use a more flexible distribution such as the Negative Binomial.

Example 2: If $Y_i \sim \text{Binomial}\left(n = 10,\ p = \frac{\beta x_i}{10}\right)$ then $\mathbb{E}(Y_i) = np = \beta x_i$ for each i.

(Note: we could use $p = \gamma x_i$ instead of $p = \frac{\beta x_i}{10}$. We use $\frac{\beta x_i}{10}$ so that we can compare the distribution of Y_i between Examples 1, 2, and 3 with the same value of β.)

This model could be suitable if Y_i measures a ***score out of 10*** on some test. In the school example, we might expect older children (larger x_i) to achieve a higher score than younger children on the same test.

Properties of $Y_i \sim$ Binomial $\left(10, \frac{\beta x_i}{10}\right)$:



- Y_i takes values $0, 1, \ldots, 10$ with a strict upper limit at 10.
- $\text{Var}(Y_i) = np(1-p) = \beta x_i\left(1 - \frac{\beta x_i}{10}\right)$

which is respectively 2·1, 2·4, 0·9 when $\beta = 3$ and $x_i = 1, 2,$ and 3.

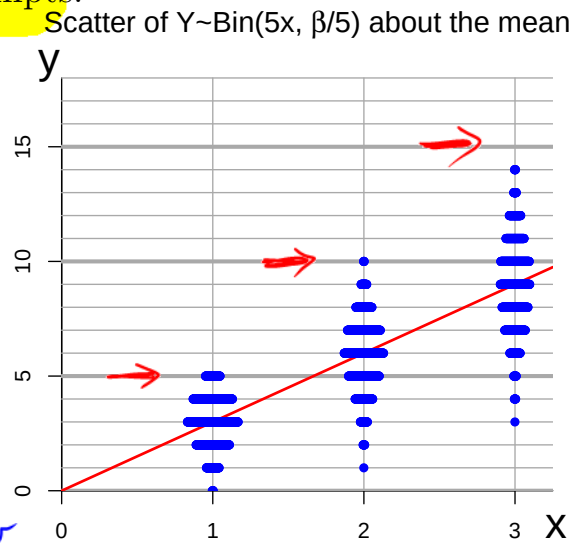
The variance becomes small as $\mathbb{E}Y$ gets close to the upper limit of 10.

Example 3: If $Y_i \sim \text{Binomial}\left(5x_i,\ p = \frac{\beta}{5}\right)$, then $\mathbb{E}(Y_i) = np = \beta x_i$ for each i.

(We use the peculiar formulation $p = \frac{\beta}{5}$ so we can keep the same value of β to compare with Examples 1 & 2.)

For example, a person at a fairground can pay \$$x$, corresponding to \$1, \$2, or \$3, to get respectively 5, 10, or 15 chances to throw a ball through a net. Their winnings are related to Y_i, the number of times they succeed in throwing the ball through the net out of their $5x_i$ attempts.

Properties of $Y_i \sim$ Binomial $\left(5x_i, \frac{\beta}{5}\right)$:



- Y_i takes values $0, 1, 2, \ldots, 5x_i$ with an upper limit at $5x_i$.
- $\text{Var}(Y_i) = np(1-p) = \beta x_i\left(1 - \frac{\beta}{5}\right)$, which is respectively 1.2, 2.4, 3.6 when $\beta = 3$ and $x = 1, 2, 3$.

The variance increases with $\mathbb{E}Y$, but (unusually) $\text{Var}(Y)$ is smaller than $\mathbb{E}Y$.

Difference between statistical regression and our previous models

- In section 3.6, we had n independent random observations $Y_1, \ldots, Y_n$. These observations were drawn from the same distribution: they were ***independent, identically distributed (iid).*** In the example in section 3.6, each $Y_i \sim \text{Poisson}(\lambda)$, and we wanted to estimate the common parameter λ.

- In statistical regression, we again have n independent random variables $Y_1, \ldots, Y_n$, but this time they have ***different distributions:*** e.g. $Y_i \sim \text{Poisson}(\beta x_i)$: distn changes as x changes. The different distributions are linked by a common parameter, β, that describes how the distribution of the response variable Y changes as the predictor variable x changes.

Estimation by maximum likelihood

"justify all steps"

To estimate the parameter β, we use maximum likelihood as usual.

We assume that the response variables $Y_1, \ldots, Y_n$ are independent, conditional on the corresponding predictor variables $x_1, \ldots, x_n$.

For observations $Y_1 = y_1, \ldots, Y_n = y_n$, the likelihood is:

$$L(\beta; y_1, \ldots, y_n) = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n \mid x_1, \ldots, x_n; \beta)$$
$$= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \mid x_i; \beta) \quad \text{by independence}$$

for $0 < \beta < \infty$.

The log-likelihood is:

$$\log\{L(\beta; y_1, \ldots, y_n)\} = \log\left\{\prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \mid x_i; \beta)\right\}$$
$$= \sum_{i=1}^{n} \log\{\mathbb{P}(Y_i = y_i \mid x_i; \beta)\}.$$

We maximize the likelihood (or more often, the log-likelihood) with respect to β as usual. The only difference from typical likelihood maximization is that we have to remember $\mathbb{P}(Y_i = y_i \mid x_i; \beta)$ is different for every different value of x_i.

Ass 3 Q3 & 4
& Task 4 on Mon

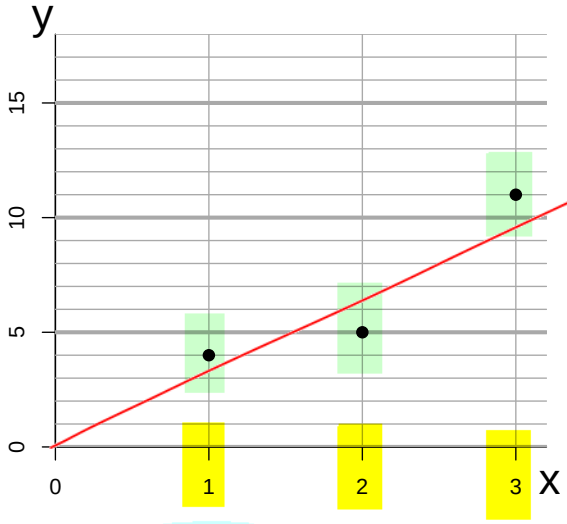
Example: Poisson regression

Recall the scenario shown at the beginning of this section:

$(x_1, y_1) = (1, 4)$

$(x_2, y_2) = (2, 5)$

$(x_3, y_3) = (3, 11)$



?? where should the line $\mathbb{E}Y = \beta x$ go? i.e. what is $\hat{\beta}$?

Consider the model $Y_i \sim \text{Poisson}(\beta x_i)$, for $i = 1, \ldots, n$.
Maximize the likelihood to find the maximum likelihood estimator, $\widehat{\beta}$.

Also find the exact variance, $\text{Var}(\widehat{\beta})$ in terms of the unknown parameter β, and suggest a suitable estimator $\widehat{\text{Var}}(\widehat{\beta})$ for the variance.

Evaluate $\widehat{\beta}$ and $\widehat{\text{Var}}(\widehat{\beta})$ for the data shown above, where $n = 3$ and $x_i = i$ for $i = 1, 2, 3$. Mark your estimated best-fit line on the graph shown.

Solution: For $Y_i \sim \text{Poisson}(\beta x_i)$, the likelihood is (from the previous page):

$$L(\beta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \mid x_i; \beta) \quad \text{(independence)}$$

$Y_i \sim \text{Poisson}(\beta x_i)$

$$= \prod_{i=1}^{n} \frac{(\beta x_i)^{y_i}}{y_i!} e^{-\beta x_i}$$

$$= \prod_{i=1}^{n} \frac{\beta^{y_i} x_i^{y_i}}{y_i!} e^{-\beta x_i}$$

$$= \underbrace{\left\{ \prod_{i=1}^{n} \frac{x_i^{y_i}}{y_i!} \right\}}_{\text{call this } K \text{ (constant)}} \beta^{\overbrace{y_1 + \cdots + y_n}^{n\bar{y}}} e^{-\beta \overbrace{(x_1 + \cdots + x_n)}^{n\bar{x}}}$$

← doesn't depend on β.

$$= K \beta^{n\bar{y}} e^{-n\bar{x}\beta} \quad \text{for } 0 < \beta < \infty.$$

Differentiate and set to 0 for MLE:

$$0 = \frac{dL}{d\beta} = \frac{d}{d\beta}\left\{ K \beta^{n\bar{y}} e^{-n\bar{x}\beta} \right\}$$

$$\Rightarrow 0 = K\left\{ n\bar{y}\, \beta^{n\bar{y}-1} e^{-n\bar{x}\beta} + \beta^{n\bar{y}}(-n\bar{x}) e^{-n\bar{x}\beta} \right\}$$

product rule.

$$0 = K n e^{-n\bar{x}\beta} \beta^{n\bar{y}-1} \left\{ \bar{y} - \bar{x}\beta \right\}$$

$$\Rightarrow \quad \beta = 0 \quad \text{or} \quad \beta = \infty \quad \text{or} \quad \bar{y} - \bar{x}\beta = 0$$

Assume a unique maximum in $0 < \beta < \infty$ ← justify by graph, or info in Q, etc.

then we must require $\bar{y} - \bar{x}\beta = 0$ at maximum,

$$\Rightarrow \hat{\beta} = \frac{\bar{y}}{\bar{x}} = \frac{\frac{1}{n}(y_1 + \dots + y_n)}{\frac{1}{n}(x_1 + \dots + x_n)} = \frac{y_1 + \dots + y_n}{x_1 + \dots + x_n}.$$

So the MLE is: $\hat{\beta} = \dfrac{Y_1 + \dots + Y_n}{x_1 + \dots + x_n}$. ML Estimator.

For data $(x_1, y_1) = (1, 4)$, $(x_2, y_2) = (2, 5)$, $(x_3, y_3) = (3, 11)$

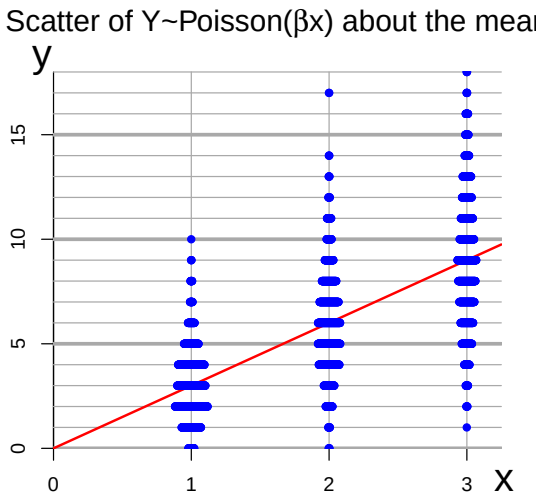
then $\hat{\beta} = \dfrac{y_1 + y_2 + y_3}{x_1 + x_2 + x_3} = \dfrac{4 + 5 + 11}{1 + 2 + 3} = \dfrac{20}{6} = 3.33.$

So $\hat{\beta} = 3.33$ is estimated slope.

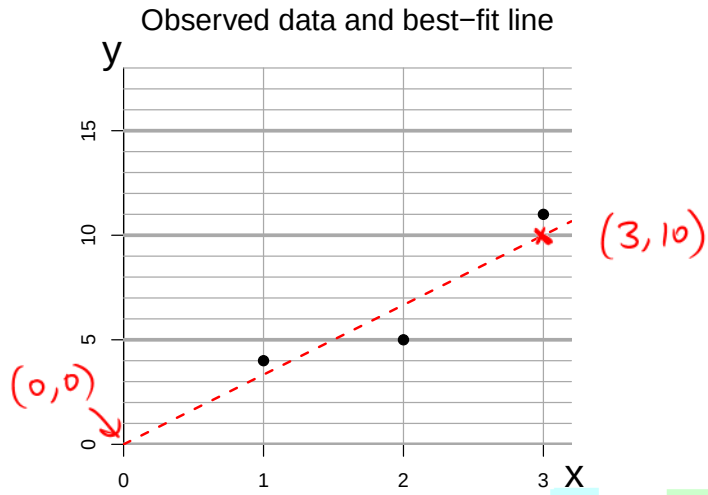
For interest, OLS would have $\hat{\beta} = 3.36$ here: close but not equal.

$\mathbb{E}Y = 3.33x$ is "best-fit line".

Add the best-fit line to the graph by picking 2 points the line passes through: $(0,0)$ and $(3, 10)$.



(a) True line with $\beta = 3$; and true distributions of $Y_i \sim \text{Poisson}(3x_i)$.



(b) Observed data and the estimated best-fit line of $y = \widehat{\beta}x$ using $\widehat{\beta} = 10/3 = 3.33$.

$Var(aX+b) = a^2 Var(X)$

Find the variance, $\text{Var}(\widehat{\beta})$, using $\widehat{\beta} = \dfrac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$: $= \dfrac{Y_1 + \dots + Y_n}{n\bar{x}}$.

So $Var(\hat{\beta}) = \dfrac{1}{(n\bar{x})^2}\{Var(Y_1 + \dots + Y_n)\}$

1 mark

$= \dfrac{1}{(n\bar{x})^2}\{Var(Y_1) + \dots + Var(Y_n)\}$ INDEP

$Y_i \sim Poisson(\beta x_i)$
so $Var(Y_i) = \beta x_i$

$= \dfrac{1}{(n\bar{x})^2}\{\beta x_1 + \beta x_2 + \dots + \beta x_n\}$

because $Y_i \sim Poisson(\beta x_i)$

$= \dfrac{1}{n^2\bar{x}^2}\beta(n\bar{x})$

$$Var(\hat{\beta}) = \frac{\beta}{n\bar{x}}.$$

Notice $Var(\hat{\beta})$ decreases as sample size n increases, if $\bar{x}$ is roughly fixed.

Suggested estimator, $\widehat{\text{Var}}(\widehat{\beta})$: $\hat{Var}(\hat{\beta}) = \dfrac{\hat{\beta}}{n\bar{x}}$.

For the data above with $\widehat{\beta} = 3.333$: $\hat{Var}(\hat{\beta}) = \dfrac{3.333}{1+2+3} = 0.556.$

$se(\hat{\beta}) = \sqrt{0.556}$, 95% approx CI is $\hat{\beta} \pm 1.96\, se(\hat{\beta})$ as usual.