Common Distributions

Table of Contents

Definitions

> Any non-negative function that sums to 1 is a proper probability function.
>
> Any non-negative function that integrates to 1 is a proper density function.

Bernoulli Distribution

> **Definition:**
> A *Bernoulli trial* is an experiment with 2 possible outcomes. The outcomes are often labelled as "success" and "failure".

If a Bernoulli trial defines the random variable

$$X = \begin{cases} 1 & \text{if the trial results in success} \\ 0 & \text{otherwise} \end{cases}$$

> Then X is said to have a Bernoulli distribution. The probability of function is:

> $$f_X(x) = \mathbb{P}(X = x) = p^x(1-p)^{1-x}$$
>
> which can also be expressed in cases form, where p when $x = 1$ and $1 - p$ when $x = 0$.

A constant like p above in a probability or density function is called a **parameter**.

Binomial Distribution

The binomial distribution arises when several independent Binomial trials are repeated in succession.

> Consider a sequence of n independent Bernoulli trials each with success probability p, if
>
> $$X = \text{total number of successes out of } n \text{ trials}$$
>
> then
>
> $$X \sim \text{Bin}(n, p)$$

As p changes, the shape of the distribution changes. As p grows larger beyond 0.5, we have a left skew distribution, and vice-versa. At $p = 0.5$, we have a normal shape.

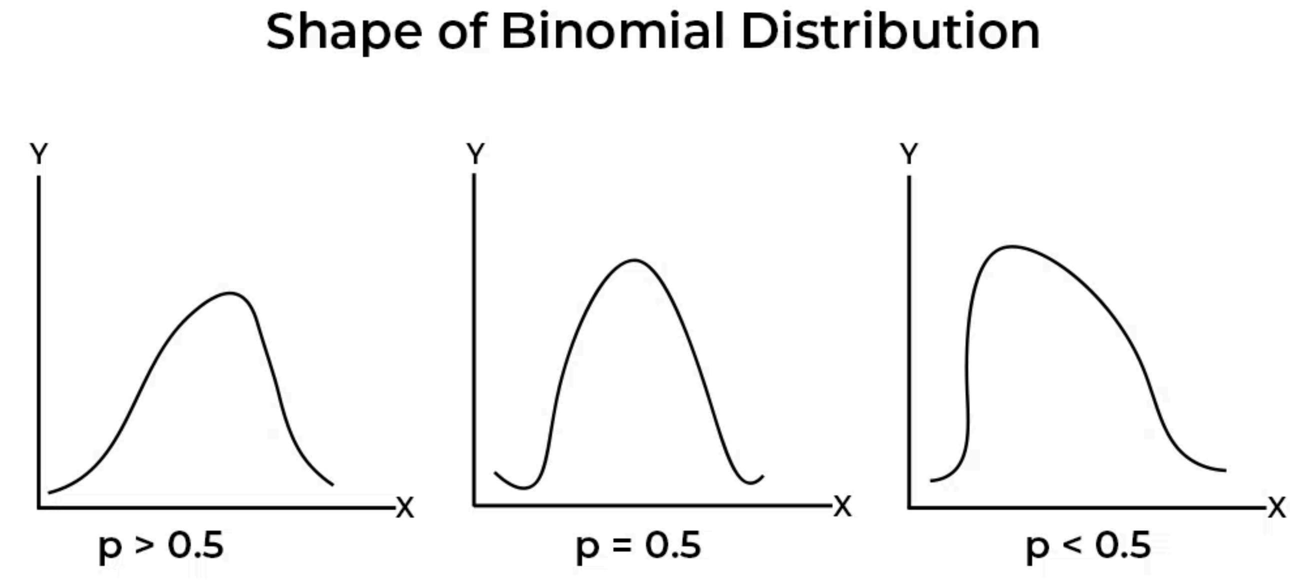
Example: The number of patients who survive a new type of surgery, out of 12 patients who each have a 95% chance of surviving.

Solution: $\text{Bin}(12, 0.95)$

Example: The number who would vote Coaliation ahead of Labor in a random sample of 365 voters (with unknown probability p of voting Coalition)

Solution: $\text{Bin}(365, p)$

> If $X \sim \text{Bin}(n, p)$ then its probability function is given by:
>
> $$f_X(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$
>
> where x is the number of occurences of some instance $x \in x$, n is the number of trials, and p is the probability of x occuring.
>
> If $X \sim \text{Bin}(n, p)$ then:
>
> 1. $E(X) = np$
> 2. $\text{Var}(X) = np(1-p)$

> X is a Bernoulli distribution if $X \sim \text{Bin}(1, p)$ (1 trial).
>
> 
>

Example: Adam pushes 48 pieces of buttered toast to test whether buttered toast more likely lands butter side up. They found that *twenty-nine* of these landed butter side up.

1. What distribution could be used to model the number of slices of buttered toast that landed buttered side up? Assume that there is a 50:50 chance of each slice landing butter side up.
2. What is the expected number of pieces of buttered toast that land butter side up, and the standard deviation?
3. What is the probability that exactly 29 slices land butter side up?
4. Calculate $\mathbb{P}(X \geq 29)$ and comment on whether this is unusual.

Solution:

1. Where X is a random variable $\in \{0, 1\}$, where 1 represents a toast landing butter side up, we have:

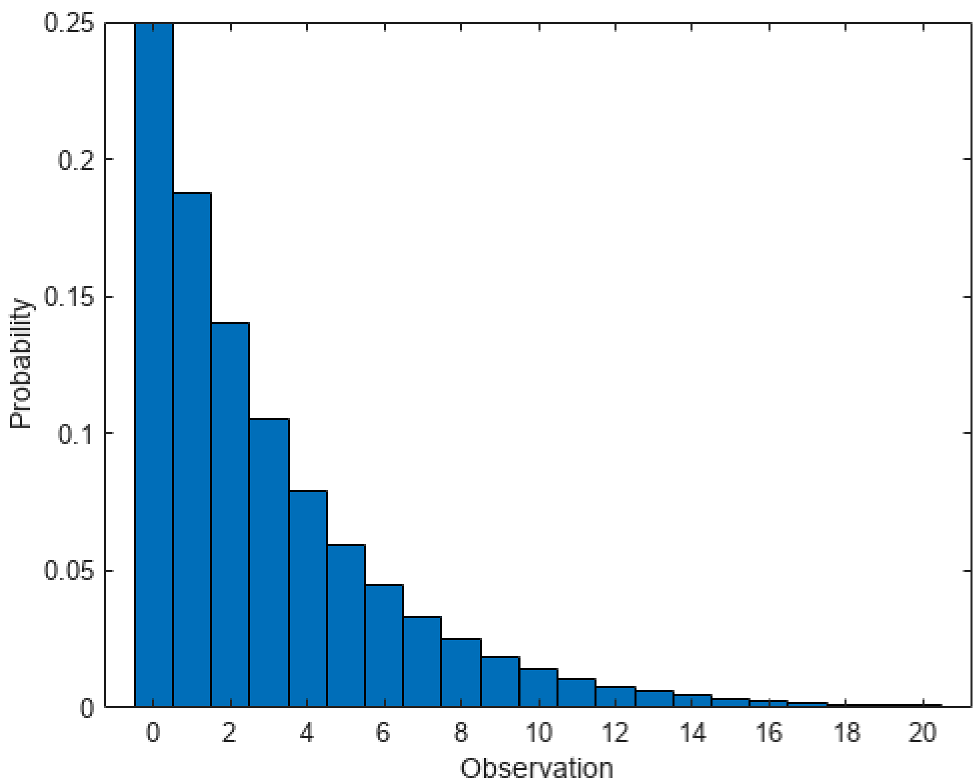
$$X \sim \text{Bin}(48, 0.5)$$

2. $\mathbb{E}(X) = 24$, $\text{std}(X) = \sqrt{12}$
3. 0.04
4. 0.097, so around 10% which is not too unusual.

Geometric Distribution

The geometric distribution arises when a Bernoulli trial is repeated until the first "success". In this case

$$X = \text{number of trials until first success}$$

> If X has a geometric distribution with parameter $0 < p < 1$, then X has a probability function
>
> $$f_X(x; p) = p(1-p)^{x-1}, x = 1, 2, 3, \ldots$$
>
> Common abbreviations are:
> $X \sim \text{Geometric}(p)$ or $X \sim \text{Geo}(p)$. $x - 1$ represents the number of failures, where p represents the first success.
>
> The geometric distribution results in:
>
> 1. $\mathbb{E}(X) = \frac{1}{p}$
> 2. $\text{Var}(X) = \frac{1-p}{p^2}$
>
> 
>

Hypergeometric Distribution

Hypogeometric random variables arise when counting the number of binary responses, when objects are sampled independently from finite populations, and the total number of "successes" in the population is known. x

Suppose that a box containts N balls, m are red and $N - m$ are black. Next, suppose that n balls are drawn at random; then:

$$X = \text{number of red balls drawn.}$$

Then X has a **hypergeometric distribution** with parameters N, m and n.

This can be thought of as a *finite population version* of the binomial distribution. Hypergeomtric distributions are considered without replacement.

> If X has a hypergeometric distribution with parameters N, m and n, then it's probability function is given by:
>
> $$f_X(x; N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$
>
> for $\max(0, n+N-m) \leq x \leq \min(m, n)$, and where $x \leq m$ and $n - x \leq N - m \implies x \geq n + N - m$. A common abbreviation is:
>
> $$X \sim \mathrm{Hg}(N, m, n)$$
>
> The following results can be derived:
>
> 1. $\mathbb{E}(X) = n \cdot \frac{m}{N}$
> 2. $\mathrm{Var}(X) = n \cdot \frac{m}{N}(1 - \frac{m}{N})(\frac{N-n}{N-1})$

Example: A lotto machine contains 45 balls and you select 6. Seven winning numbers are then drawn (6 main, one supplementary), and you win a major prize ($100000+) if you pick six of the winning numbers.

Whats the chance that you win a major prize from playing one game?

Solution:

- $N = 45$, $m = 6$, $n = 7$
- There are 45 total samples, 6 must be chosen to win, and 7 balls contain the conditions to win.
- Therefore:

$$\mathbb{P}(X = 6) = \frac{\binom{6}{6}\binom{39}{1}}{\binom{45}{7}}$$

which equals 0.00000086.

Poisson Distribution

> The random variable X has a **Poisson** distribution with parameter $\lambda > 0$ if its probability function is

$$f_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$$

Common abbreviations are:

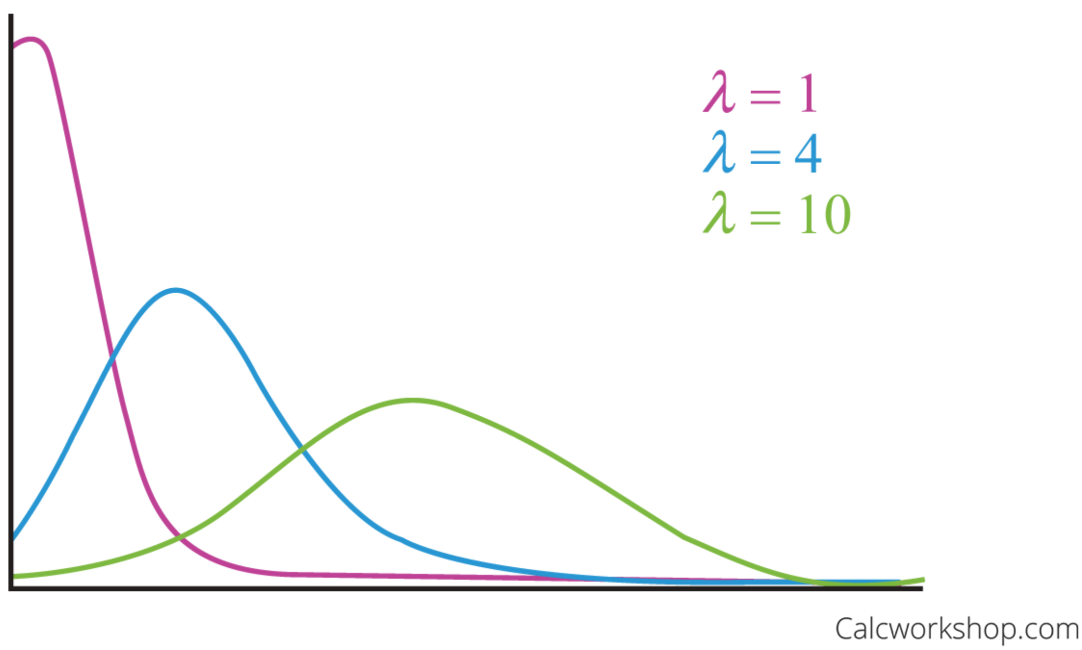
$$X \sim \text{Poisson}(\lambda) \text{ or } \text{Poi}(\lambda)$$

The following results can be derived:

1. $\mathbb{E}(X) = \lambda$
2. $\text{Var}(X) = \lambda$

Some general features are:

- Centred around λ, with spread of λ
- As $\lambda \to \infty$, the distribution shifts over from right skew to left skew. ("Rare" events become more oftenly seen)



The parameter λ is often called the intensity, rate or event rate parameter. The Poisson distribution often arises when the **variable of interest is a count**. For example, the number of traffic accdients in a city on any given day could be well-described by a Poisson random variable.

The Poisson distribution is also a standard distribution for the occurence of *rare events*. Such events are often described by a *Poisson process*.

- A Poisson process is a model for the occurence of point events in a continuum, usually a time-continuum but can also be spatial.

- The occurence of points in disjoint intervals is independent, with a uniform probability rate over time
- If the occurence rate is λ, then the number of points ocurring in a time interval of length t is a random variable with a $\text{Poisson}(\lambda t)$.

Example: Model the number of workplace accidents in a month (when the average number of accidents per month is 1.4)

Solution: $\text{Poi}(1.4)$; we haven't been given an n; and we've been given some events frequency that is a non-normal event.

Example: If, on average, five university servers go offline per week, what is the chance that no more than one will go offline this week? (assume independence)

Solution: Let X be the number of servers that go offline this week. We're told that on average 5 servers go offline; so we have the distribution $X \sim \text{Poi}(5)$.

We want to find $P(X \leq 1)$, and therefore we have $P(X = 0) + P(X = 1)$, therefore we have:

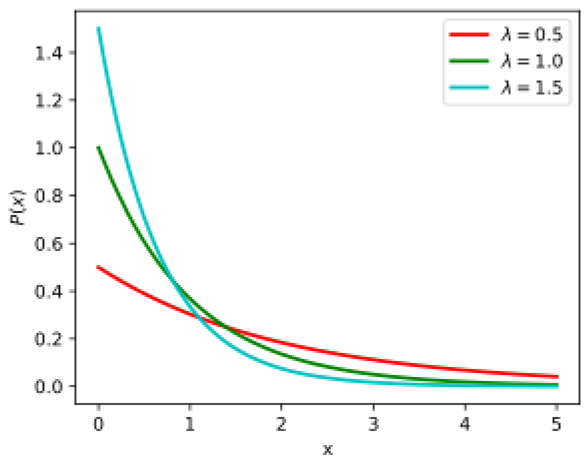
$$\frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!}$$
$$\approx 0.0404$$

This can be done on RStudio using `ppois(1, 5)`, which finds the cumulative probability from `x = 0 -> 1`. Distributions pre-pended with `p` find $P(X \leq a)$ probabilities.

Exponential Distributions

> A random variable X is said to have an **exponential distribution** with parameter $\beta > 0$, if X has density function
>
> $$f_X(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, x > 0$$
>
> Common abbreviations are:
>
> $$X \sim \text{Exponential}(\beta) \text{ or } \text{Exp}(\beta)$$
>
> A parameterisation of the exponential distribution is:
>
> $$f_X(x; \lambda) = \lambda e^{-\lambda x}, x > 0$$
>
> where $\lambda = 1/\beta$.
>
> The following results can be derived:

1. $\mathbb{E}(X) = \beta$
2. $\text{Var}(X) = \beta^2$



Exponential distributions describe the probability structure of *positive* random variables, such as modelling the lifetime of batteries, or waiting times until the next earthquake.

We can interpret the parameters of the exponential distribution as:

- λ, which represents the average rate at which events occur over time
- β is the mean time between events.

Example: If, on average, 5 servers go offline during the week, what is the chance that no servers will go offline today? (Note that a day is one-seventh of a week)

Solution: Let T be the waiting time in weeks between servers going offline, therefore $T \sim \text{Exp}(\beta)$. $B = 1/\lambda$, and since $\lambda = 5$, $\beta = 1/5$. Thus, we have $T = 1/7$, since we want to find the probability that no servers go offline today, giving:

$$f_T(\frac{1}{7}; \frac{1}{5}) = 5e^{-(1/7)\cdot 5}$$
$$\approx 0.49$$

Lack of memory

An important property of the exponential distribution is *lack of memory*; if X has an exponential distribution, then:

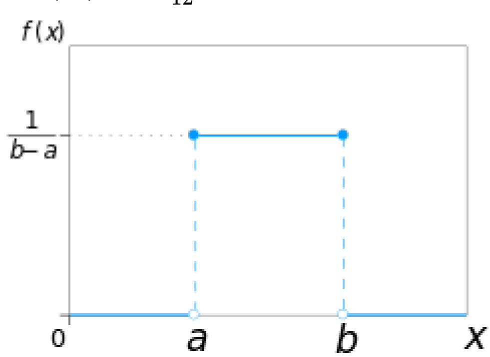
$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)$$

In words, if the waiting time until the next event is exponential, then the next event is indenpendent of the time you've already been waiting.

Proof:

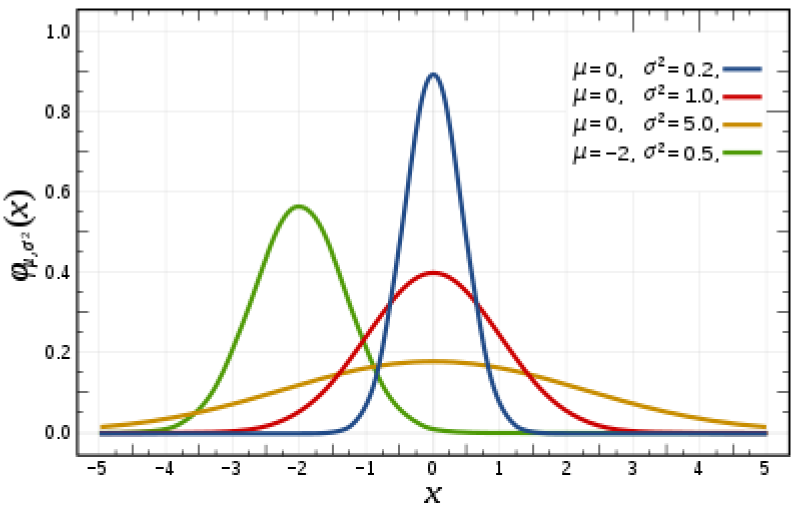
$$\begin{aligned}
\mathbb{P}(X > s+t | X > s) &= \frac{P((X > s+t) \cap (X > s))}{P(X > s)} \\
&= \frac{P(X > s+t)}{P(x > s)} \\
&= \frac{e^{(-s+t)/\beta}}{e^{-s/\beta}} \\
&= e^{-t/\beta} \\
&= \mathbb{P}(X > t)
\end{aligned}$$

Uniform distribution

> A continuous random variable X that can take values in the interval (a, b) with equal likely is said to have a **uniform** distribution on (a, b).
> Common abbreviations are:
>
> $$X \sim \text{Uniform}(a, b) \text{ or } X \sim \text{Unif}(a, b)$$
>
> If $X \sim \text{Uniform}(a, b)$ then the density function of X is
>
> $$f_X(x; a, b) = \frac{1}{b-a}$$
>
> for $a < x < b$ and $a < b$.
> The following results can be derived:
>
> 1. $\mathbb{E}(X) = \frac{a+b}{2}$
> 2. $\text{Var}(X) = \frac{(b-a)^2}{12}$
>
> 
>

Basically just a rectangle - and denotes that for some range (a, b) that every event has the same probability.

Normal Distribution

> A continuous random variable X is said to have a **normal distribution** or a **Gaussian distribution** with parameters μ and σ^2 (where $-\infty < \mu < \infty$ and $\sigma^2 > 0$) if X has density function:
>
> $$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
>
> A common shorthand is
>
> $$X \sim N(\mu, \sigma)^2$$
>
> The following results can be derived:
>
> 1. $\mathbb{E}(X) = \mu$
> 2. $\text{Var}(X) = \sigma^2$
>
> 
>

The standard normal distribution expressed by Z, we have:

$$Z \sim N(0, 1)$$

Or mathematically expressed by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The standard normal distribution does not have a closed form anti-derivative, and thus requires `RStudio` to calculate the probability. More commonly, we use the notion of z-scores to calculate the probabilities; given by:

$$z = \frac{x - \mu}{\sigma}$$

These z scores can be used within `pnorm` to find the probability of any x values in any N distribution by standardisation.

> If $Z \sim N(0, 1)$ then

> $$P(Z \leq x) = F_Z(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \phi(x)$$
>
> where $\phi(x)$ is used to denote the **cumulative distribution function** of the $N(0,1)$

With a reminder that `pnorm` or any distribution function in `RStudio` that is prepended with `p` denotes a `CDF`, we use `pnorm(0.47)` to find the probability of $\phi(0.47)$. The function `qnorm` finds the quantiles x from the standard normal for some given probability.

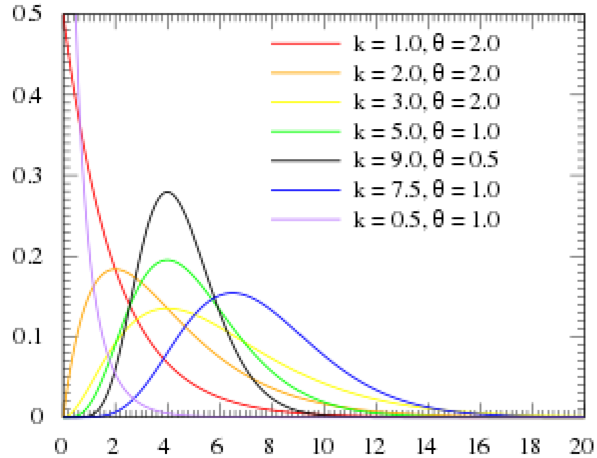
$\phi(x)$ limits to 0 on the negative axis, 1 on the positive axis, $\phi(0) = 1/2$ and ϕ is monotonically increasing.

Gamma Distribution

Gamma distributions is useful for skewed data, and like the exponential distribution, it's support is always posiitve.

> The **gamma function** at $x \in \mathbb{R}$ with parameter t is given by:
>
> $$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$
>
> Some results for the gamma function are:
>
> 1. $\Gamma(x) = (x-1)\Gamma(x-1)$
> 2. $\Gamma(n) = (n-1)!$
> 3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

> A random variable X is said to have a **gamma distribution** with parameters α and β (where $\alpha, \beta > 0$) if X has density function
>
> $$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}$$
>
> - The parameter α is known as the shape parameter, and it's role is to change the shape of the curve.
> - The parameter β is known as the scale parameter, it's role is to stretch or shrink the curve
>
> A common notation for this distribution is:
>
> $$X \sim \text{Gamma}(\alpha, \beta)$$
>
> The following results can be derived:

> 1. $\mathbb{E}(X) = \alpha\beta$
> 2. $\text{Var}(X) = \alpha\beta^2$
>
> 
>

Gamma distributions, akin to normal distributions, have some half-line; and then are skewed by α. As α increases, we focus more into a half line. β changes the spread from the half line.

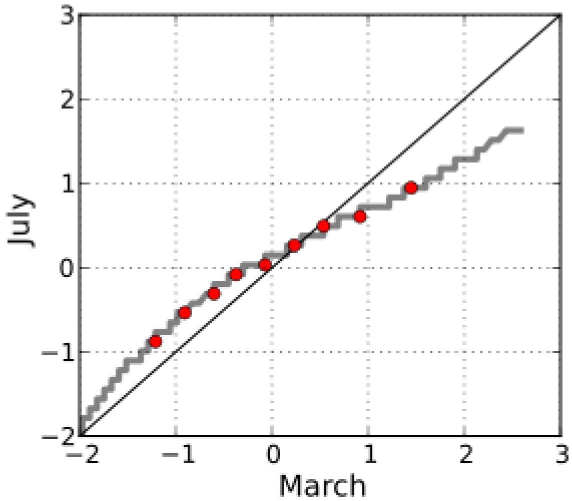
> X has an exponential distribution if and only if
>
> $$X \sim \text{Gamma}(1, \beta)$$

Quantile-quantile plots

Consider the situation in which we have a sample of size n from some unknown random variable $x_1, x_2, \dots$ and we want to check if the data appears to come from a random variable with cdf $F_X(x)$

This can be achieved using a **quantile-quantile plot** (sometimes called a Q-Q plot) as defined below:

> **Q-Q plots**: To check how well the sample X_s approximates the distribution with cdf $F_X(x), plotthe\$n$ sample quantiles against the corresponding quantiles of $F_X(x)$:
>
> $$F^{-1}(p, x_(k))$$
>
> where $p = \frac{k-0.5}{n}$ for all $k = \{1, 2, 3, \dots\}$



> The above plot approximates closely to the theoretical quantiles, and thus is likely of the distribution.

RStudio

Consider the normal distribution `norm`

To take a **random sample** of size 20 from X and store it in an object called X, we do:

> `x <- rnorm(20, 2, 3)`

To find the **density function** of X at the value $x = 4$, we use:

> `dnorm(4, 2, 3)`

To find the **probability/CDF** of $F_X(4)$, we use:

> `pnorm(4, 2, 3)`

To find the 95-th percentile of X, denote as x_p where $p = 0.95$, we use:

> `qnorm(0.95, 2, 3)`

To plot the Q-Q plot between a random gamma sample, and a normal distribution

```
x <- rgamma(1000, 1, 10)
qqnorm(x) # this compares to normal distribution
qqline(x)
```