0 Types of distributions

A distribution in statistics is a function that shows the possible values for a variable Y and how often they occur.

Think about a die. It has six sides, numbered from 1 to 6. We roll the die. What is the probability of getting 1? It is one out of six, so one-sixth, right? What is the probability of getting 2? Once again one-sixth. The same holds for 3, 4, 5 and 6.

Now. What is the probability of getting a 7? It is impossible to get a 7 when rolling a die. Therefore, the probability is 0.

The distribution of an event consists not only of the input values that can be observed, but is made up of all possible values.

So, the distribution of the event - rolling a die - will be given by the following: The probability of getting one is 0.16666, the probability of getting 2 is 0.16666, and so on. You are sure that you have exhausted all possible values when the sum of probabilities is equal to 1 or 100%. For all other values, the probability of occurrence is 0.

Each probability distribution is associated with a graph describing the likelihood of occurrence of every event. It is crucial to understand that the distribution in statistics is defined by the underlying probabilities and not the graph. The graph is just a visual representation.

The type of distribution to use depends on the type of variable being modelled. We can consider three distinct types of distributions:

1. Continuous distributions; see Figure 1(a),
2. Discrete distributions; see Figure 1(a),
3. Mixed distributions; see Figure 1(c),

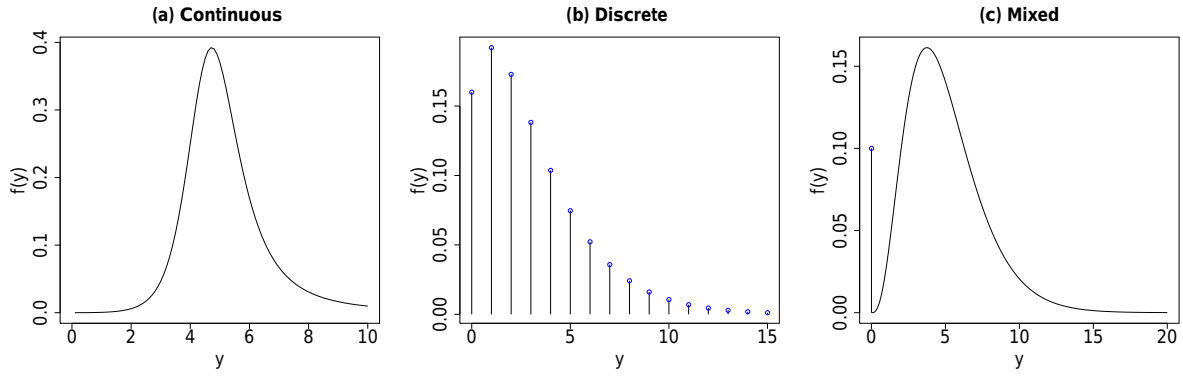

Figure 1: Different type of distributions: (a) continuous, (b) discrete, (c) mixed.

We will be looking at some of the continuous and discrete distributions here.

Distribution	Range
Beta	$(0, 1)$
Dagum	$(0, +\infty)$
Log-logistic	$(0, +\infty)$
Gamma	$(0, +\infty)$
Log-normal	$(0, +\infty)$
Singh-Maddala	$(0, +\infty)$
Weibull	$(0, +\infty)$
Gumbel	$(-\infty, +\infty)$
Reverse Gumbel	$(-\infty, +\infty)$
Logistic	$(-\infty, +\infty)$
Normal	$(-\infty, +\infty)$

Table 1: Some continuous distributions.

0.1 Continuous distributions

Continuous distributions are usually defined on $(-\infty, +\infty)$, $(0, +\infty)$ or $(0, 1)$, but can take other support ranges. Table 1 shows some these distributions.

0.1.1 Beta

- $0 < y < 1$
- $Y \sim \text{Beta}(\mu, \sigma)$, $0 < \mu < 1$, $0 < \sigma < 1$
- $F(y|\mu, \sigma) = I(y; \alpha_1, \alpha_2)$
- $f(y|\mu, \sigma) = \frac{y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)}$
- $\alpha_1 = \frac{\mu(1-\sigma^2)}{\sigma^2}$ and $\alpha_2 = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$
- $I(\cdot; \cdot, \cdot)$ is the regularized beta function
- $B(\cdot, \cdot)$ is the beta function
- $E(Y) = \mu$
- $V(Y) = \sigma^2\mu(1-\mu)$

This distribution is useful for situations where the variable of interest is continuous and restricted to the interval (0, 1). So is used to model rates and proportions. Some example are: proportion of household income spent on food; data on a cross-section of countries in which inequality is measured using the Gini coefficient. The Gini coefficient is one if one person has all the income in a society and zero if income is equally divided among everyone. Values of zero and one simply do not happen, of course.

0.1.2 Dagum

- $y > 0$
- $Y \sim \text{Dagum}(\mu, \sigma, \nu)$, $\mu > 0$, $\sigma > 0$, $\nu > 0$
- $F(y|\mu, \sigma, \nu) = \left\{1 + \left(\frac{y}{\mu}\right)^{-\sigma}\right\}^{-\nu}$
- $f(y|\mu, \sigma, \nu) = \frac{\sigma\nu}{y}\left[\frac{\left(\frac{y}{\mu}\right)^{\sigma\nu}}{\left\{\left(\frac{y}{\mu}\right)^{\sigma}+1\right\}^{\nu+1}}\right]$
- $E(Y) = -\frac{\mu}{\sigma}\frac{\Gamma\left(-\frac{1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}$ if $\sigma > 1$
- $V(Y) = -\left(\frac{\mu}{\sigma}\right)^2\left[2\sigma\frac{\Gamma\left(-\frac{2}{\sigma}\right)\Gamma\left(\frac{2}{\sigma}+\nu\right)}{\Gamma(\nu)} + \left\{\frac{\Gamma\left(-\frac{1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}\right\}^2\right]$ if $\sigma > 2$

Note that the Dagum distribution is derived from a more complex distribution, the generalized beta type 2, characterized by four parameters: μ, σ, ν and τ. When $\tau = 1$, then we have the Dagum distribution. When both τ and ν are equal to 1, we have the log-logistic distribution.

The Dagum distribution is a continuous probability distribution defined over positive real numbers. It is mostly associated with the study of income distribution. In economics, income distribution is how a nation's total GDP is distributed amongst its population.

0.1.3 Gamma

- $y > 0$
- $Y \sim \text{Gamma}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $F(y|\mu, \sigma) = \frac{1}{\Gamma\left(\frac{1}{\sigma^2}\right)}\gamma\left(\frac{1}{\sigma^2}, \frac{y}{\mu\sigma^2}\right)$
- $f(y|\mu, \sigma) = \frac{1}{(\mu\sigma^2)^{\frac{1}{\sigma^2}}}\frac{y^{\frac{1}{\sigma^2}-1}\exp\left(-\frac{y}{\mu\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)}$
- $\Gamma(\cdot)$ is the gamma function
- $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function
- $E(Y) = \mu$
- $Va(Y) = \mu^2\sigma^2$

The Gamma distribution is a continuous probability distribution defined over positive real numbers. In the health economic literature there are many examples of modelling health cost data using a gamma distribution.

0.1.4 Log-Normal

- $y > 0$
- $Y \sim \text{Log-Normal}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $F(y|\mu,\sigma) = \frac{1}{2} + \frac{1}{2}\text{erf}\left\{\frac{\log(y)-\mu}{\sigma\sqrt{2}}\right\}$
- $f(y|\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$
- $\text{erf}(\cdot)$ is the error function
- $E(Y) = \sqrt{\exp\left(\sigma^2\right)}\exp\left(\mu\right)$
- $V(Y) = \exp\left(\sigma^2\right)\left\{\exp\left(\sigma^2\right) - 1\right\}\exp\left(2\mu\right)$

There is a nice relationship with the normal distribution. If $\widetilde{Y} \sim \text{Log-Normal}(\mu, \sigma)$ then $Y = \log(\widetilde{Y}) \sim \mathcal{N}(\mu, \sigma)$; likewise, if $Y \sim \mathcal{N}(\mu, \sigma)$ then $\widetilde{Y} = \exp(Y) \sim \text{Log-Normal}(\mu, \sigma)$.

0.1.5 Singh-Maddala

- $y > 0$
- $Y \sim \text{Singh-Maddala}(\mu, \sigma, \nu)$, $\mu > 0$, $\sigma > 0$, $\nu > 0$
- $F(y|\mu,\sigma,\nu) = 1 - \left\{1 + \left(\frac{y}{\mu}\right)^\sigma\right\}^{-\nu}$
- $f(y|\mu,\sigma,\nu) = \frac{\sigma\nu y^{\sigma-1}}{\mu^\sigma\left\{1+\left(\frac{y}{\mu}\right)^\sigma\right\}^{\nu+1}}$
- $E = \mu\frac{\Gamma\left(1+\frac{1}{\sigma}\right)\Gamma\left(-\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}$ if $\sigma\nu > 1$
- $V(Y) = \mu^2\left\{\Gamma\left(1+\frac{2}{\sigma}\right)\Gamma(\nu)\Gamma\left(-\frac{2}{\sigma}+\nu\right) - \Gamma\left(1+\frac{1}{\sigma}\right)^2\Gamma\left(-\frac{1}{\sigma}+\nu\right)^2\right\}$ if $\sigma\nu > 2$

Note that the Singh-Maddala distribution is derived from the generalized beta type 2. When ν is equal to 1, we have the Singh-Maddala distribution.

0.1.6 Weibull

- $y > 0$
- $Y \sim \text{Weibull}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $F(y|\mu,\sigma) = 1 - \exp\left\{-\left(\frac{y}{\mu}\right)^\sigma\right\}$
- $f(y|\mu,\sigma) = \frac{\sigma}{\mu}\left(\frac{y}{\mu}\right)^{\sigma-1}\exp\left\{-\left(\frac{y}{\mu}\right)^\sigma\right\}$
- $E(Y) = \mu\Gamma\left(\frac{1}{\sigma}+1\right)$
- $V(Y) = \mu^2\left[\Gamma\left(\frac{2}{\sigma}+1\right) - \left\{\Gamma\left(\frac{1}{\sigma}+1\right)\right\}^2\right]$

This distribution can be used with survival data.

0.1.7 Gumbel

- $-\infty < y < \infty$
- $Y \sim \text{Gumbel}(\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$
- $F(y|\mu, \sigma) = 1 - \exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$
- $f(y|\mu, \sigma) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}$
- $E(Y) = \mu - 0.57722\sigma$
- $V(Y) = \frac{\pi^2\sigma^2}{6}$

The Gumbel distribution is a continuous probability distribution defined over real numbers. It is an asymmetric distribution and often is used in extreme value analysis.

0.1.8 Reverse Gumbel

- $-\infty < y < \infty$
- $Y \sim \text{Reverse Gumbel}(\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$
- $F(y|\mu, \sigma) = \exp\left\{-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}$
- $f(y|\mu, \sigma) = \frac{1}{\sigma}\exp\left\{\left(-\frac{y-\mu}{\sigma}\right) - \exp\left(-\frac{y-\mu}{\sigma}\right)\right\}$
- $E(Y) = \mu + 0.57722\sigma$
- $V(Y) = \frac{\pi^2\sigma^2}{6}$

There is a nice relationship with the Gumbel distribution. If $Y \sim \text{Reverse Gumbel}(\mu, \sigma)$ and $\widetilde{Y} = -Y$ then $\widetilde{Y} \sim \text{Gumbel}(-\mu, \sigma)$.

0.1.9 Logistic

- $-\infty < y < \infty$
- $Y \sim \text{Logistic}(\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$
- $F(y|\mu, \sigma) = \frac{1}{1+\exp\left(-\frac{y-\mu}{\sigma}\right)}$
- $f(y|\mu, \sigma) = \frac{1}{\sigma}\left\{\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}\left\{1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right\}^{-2}$
- $E(Y) = \mu$
- $V(Y) = \frac{\pi^2\sigma^2}{3}$

0.1.10 Normal

- $-\infty < y < \infty$
- $Y \sim \mathcal{N}(\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$
- $F(y|\mu, \sigma) = \frac{1}{2}\left\{1 + \text{erf}\left(\frac{y-\mu}{\sigma\sqrt{2}}\right)\right\}$
- $f(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$
- $E(Y) = \mu$
- $V(Y) = \sigma^2$

Both logistic and normal distribution are symmetric (bell shaped). The main difference between the normal distribution and the logistic distribution lies in the tails. The logistic distribution has slightly longer tails compared to the normal distribution.

Distribution	Range
Poisson	{0,1,2,...}
Negative binomial type I	{0,1,2,...}
Negative binomial type II	{0,1,2,...}
Poisson inverse Gaussian	{0,1,2,...}
Bernoulli	{0,1}

Table 2: Some discrete distributions.

0.2 Discrete distributions

Discrete distributions are usually defined on $y = 0, 1, 2, \ldots, n$, where n is either a known finite value, or infinite. Table 2 shows some discrete distributions.

0.2.1 Poisson

- $y \in \{(0), 1, 2, 3, \ldots\}$
- $Y \sim \text{Poisson}(\mu)$, $\mu > 0$
- $f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$
- $E(Y) = \mu$
- $V(Y) = \mu$

The crucial point in the Poisson distribution is that the mean is the same as variance. Sometimes this might be too restrictive. Can you think of some real data examples? Number of trades in a time interval; number of a given disaster i.e., default- per month; number of crimes on campus per semester.

0.2.2 Negative binomial type I

- $y \in \{0, 1, 2, 3, \ldots\}$
- $Y \sim \text{NBI}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $f(y|\mu, \sigma) = \frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$
- $E(Y) = \mu$
- $V(Y) = \mu + \sigma\mu^2$

This distribution is more flexible than the Poisson as it accounts for over-dispersion. This distribution counts the number of trials up to and including the k-th success.

0.2.3 Negative binomial type II

- $y \in \{0, 1, 2, 3, \ldots\}$
- $Y \sim \text{NBII}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $f(y|\mu, \sigma) = \frac{\Gamma(y+\mu/\sigma)\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1+\sigma)^{y+\mu/\sigma}}$
- $E(Y) = \mu$
- $V(Y) = (1 + \sigma)\mu$

This distribution accounts for overdispersion and counts the failures before the k-th success.

What is the differences between type I and type II negative binomial distribution?

We have already said that type I counts the number of trials up to and including the k-th success whereas type II counts the failures before the k-th success.

So suppose we are counting the number of goals we score (success) for the penalty kicks we make (trials).

Suppose we score our 3rd goal on our 10th penalty kick.

Under type I $Y = 10$ as it is the 10th kick when we got our 3rd goal.

Under type II $Y = 7$ as we had 7 failures before we got our 3rd goal.

0.2.4 Poisson inverse Gaussian

- $y \in \{0, 1, 2, 3, \ldots\}$
- $Y \sim \text{PIG}(\mu, \sigma)$, $\mu > 0$, $\sigma > 0$
- $f(y|\mu, \sigma) = \left(\frac{2\alpha}{\pi}\right)^{0.5} \frac{\mu^y \exp(1/\sigma) K_{y-0.5}(\alpha)}{(\alpha\sigma)^y y!}$
- $E(Y) = \mu$
- $V(Y) = \mu + \sigma\mu^2$

It is used for modelling highly dispersed count data and is an alternative to negative binomial distributions.

0.2.5 Bernoulli

- $y \in 0, 1$
- $Y \sim \text{Bernoulli}(\mu,)$, $0 < \mu < 1$
- $f(y|\mu) = \mu$ if $y = 1$,
  $f(y|\mu) = 1 - \mu$ if $y = 0$
- $E(Y) = \mu$
- $V(Y) = \mu(1 - \mu)$

0.3 Power law distributions

Material for these distributions can be find here:

1. `https://arxiv.org/abs/0706.1062`
2. `https://www.jstatsoft.org/article/view/v064i02`
3. `https://cran.r-project.org/package=poweRlaw/vignettes/c_comparing_distributions.pdf`