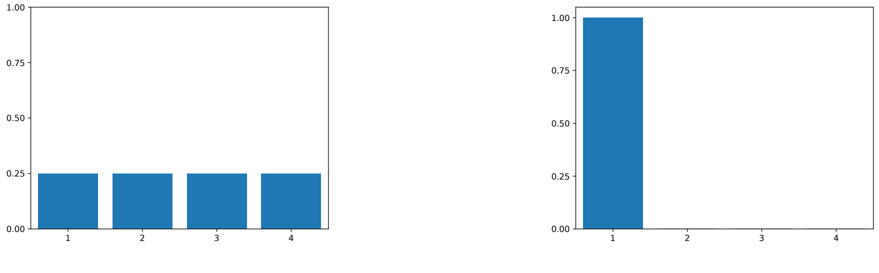


(a) *(b)*

Figure 2.1: Some discrete distributions on the state space $\mathcal{X} = \{1, 2, 3, 4\}$. (a) A uniform distribution with $p(x = k) = 1/4$. (b) A degenerate distribution (delta function) that puts all its mass on $x = 1$. Generated by code at figures.probml.ai/book1/2.1.

probability mass function or **pmf** as a function which computes the probability of events which correspond to setting the rv to each possible value:

$$p(x) \triangleq \Pr(X = x) \tag{2.8}$$

The pmf satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

If X has a finite number of values, say K, the pmf can be represented as a list of K numbers, which we can plot as a histogram. For example, Figure 2.1 shows two pmf's defined on $\mathcal{X} = \{1, 2, 3, 4\}$. On the left we have a uniform distribution, $p(x) = 1/4$, and on the right, we have a degenerate distribution, $p(x) = \mathbb{I}(x = 1)$, where $\mathbb{I}()$ is the binary indicator function. Thus the distribution in Figure 2.1(b) represents the fact that X is always equal to the value 1. (Thus we see that random variables can also be constant.)

2.2.2 Continuous random variables

If $X \in \mathbb{R}$ is a real-valued quantity, it is called a **continuous random variable**. In this case, we can no longer create a finite (or countable) set of distinct possible values it can take on. However, there are a countable number of *intervals* which we can partition the real line into. If we associate events with X being in each one of these intervals, we can use the methods discussed above for discrete random variables. By allowing the size of the intervals to shrink to zero, we can represent the probability of X taking on a specific real value, as we show below.

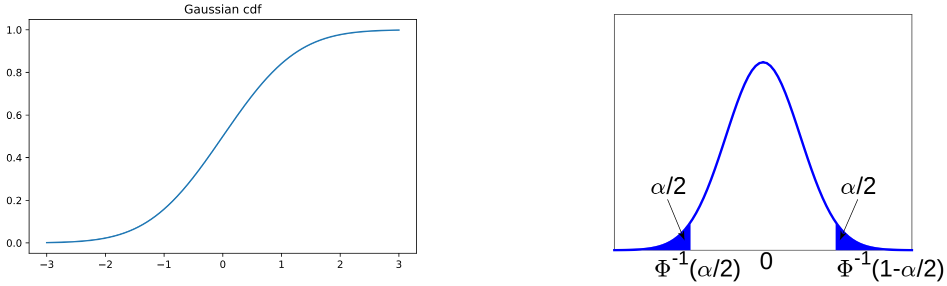
2.2.2.1 Cumulative distribution function (cdf)

Define the events $A = (X \leq a)$, $B = (X \leq b)$ and $C = (a < X \leq b)$, where $a < b$. We have that $B = A \vee C$, and since A and C are mutually exclusive, the sum rules gives

$$\Pr(B) = \Pr(A) + \Pr(C) \tag{2.9}$$

and hence the probability of being in interval C is given by

$$\Pr(C) = \Pr(B) - \Pr(A) \tag{2.10}$$



(a) *(b)*

Figure 2.2: (a) Plot of the cdf for the standard normal, $\mathcal{N}(0, 1)$. Generated by code at figures.probml.ai/book1/2.2. (b) Corresponding pdf. The shaded regions each contain $\alpha/2$ of the probability mass. Therefore the nonshaded region contains $1 - \alpha$ of the probability mass. The leftmost cutoff point is $\Phi^{-1}(\alpha/2)$, where Φ is the cdf of the Gaussian. By symmetry, the rightmost cutoff point is $\Phi^{-1}(1 - \alpha/2) = -\Phi^{-1}(\alpha/2)$. Generated by code at figures.probml.ai/book1/2.2.

In general, we define the **cumulative distribution function** or **cdf** of the rv X as follows:

$$P(x) \triangleq \Pr(X \leq x) \tag{2.11}$$

(Note that we use a capital P to represent the cdf.) Using this, we can compute the probability of being in any interval as follows:

$$\Pr(a < X \leq b) = P(b) - P(a) \tag{2.12}$$

Cdf's are monotonically non-decreasing functions. See Figure 2.2a for an example, where we illustrate the cdf of a standard normal distribution, $\mathcal{N}(x|0, 1)$; see Section 2.6 for details.

2.2.2.2 Probability density function (pdf)

We define the **probability density function** or **pdf** as the derivative of the cdf:

$$p(x) \triangleq \frac{d}{dx} P(x) \tag{2.13}$$

(Note that this derivative does not always exist, in which case the pdf is not defined.) See Figure 2.2b for an example, where we illustrate the pdf of a univariate Gaussian (see Section 2.6 for details).

Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

$$\Pr(a < X \leq b) = \int_a^b p(x)dx = P(b) - P(a) \tag{2.14}$$

As the size of the interval gets smaller, we can write

$$\Pr(x \leq X \leq x + dx) \approx p(x)dx \tag{2.15}$$

Intuitively, this says the probability of X being in a small interval around x is the density at x times the width of the interval.

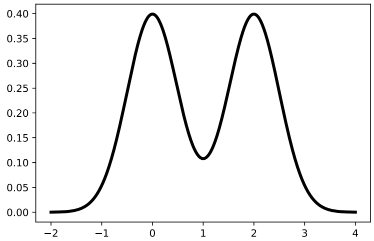


Figure 2.4: Illustration of a mixture of two 1d Gaussians, $p(x) = 0.5\mathcal{N}(x|0, 0.5) + 0.5\mathcal{N}(x|2, 0.5)$. *Generated by code at figures.probml.ai/book1/2.4.*

To prove this, let us suppose, for simplicity, that X and Y are both discrete rv's. Then we have

$$\mathbb{E}_Y\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}_Y\left[\sum_x x\, p(X = x|Y)\right] \tag{2.42}$$

$$= \sum_y \left[\sum_x x\, p(X = x|Y)\right] p(Y = y) = \sum_{x,y} xp(X = x, Y = y) = \mathbb{E}\left[X\right] \tag{2.43}$$

To give a more intuitive explanation, consider the following simple example.[2] Let X be the lifetime duration of a lightbulb, and let Y be the factory the lightbulb was produced in. Suppose $\mathbb{E}\left[X|Y = 1\right] = 5000$ and $\mathbb{E}\left[X|Y = 2\right] = 4000$, indicating that factory 1 produces longer lasting bulbs. Suppose factory 1 supplies 60% of the lightbulbs, so $p(Y = 1) = 0.6$ and $p(Y = 2) = 0.4$. Then the expected duration of a random lightbulb is given by

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[X|Y = 1\right] p(Y = 1) + \mathbb{E}\left[X|Y = 2\right] p(Y = 2) = 5000 \times 0.6 + 4000 \times 0.4 = 4600 \tag{2.44}$$

There is a similar formula for the variance. In particular, the **law of total variance**, also called the **conditional variance formula**, tells us that

$$\mathbb{V}\left[X\right] = \mathbb{E}_Y\left[\mathbb{V}\left[X|Y\right]\right] + \mathbb{V}_Y\left[\mathbb{E}\left[X|Y\right]\right] \tag{2.45}$$

To see this, let us define the conditional moments, $\mu_{X|Y} = \mathbb{E}\left[X|Y\right]$, $s_{X|Y} = \mathbb{E}\left[X^2|Y\right]$, and $\sigma^2_{X|Y} = \mathbb{V}\left[X|Y\right] = s_{X|Y} - \mu^2_{X|Y}$, which are functions of Y (and therefore are random quantities). Then we have

$$\mathbb{V}\left[X\right] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \mathbb{E}_Y\left[s_{X|Y}\right] - \left(\mathbb{E}_Y\left[\mu_{X|Y}\right]\right)^2 \tag{2.46}$$

$$= \mathbb{E}_Y\left[\sigma^2_{X|Y}\right] + \mathbb{E}_Y\left[\mu^2_{X|Y}\right] - \left(\mathbb{E}_Y\left[\mu_{X|Y}\right]\right)^2 \tag{2.47}$$

$$= E_Y[\mathbb{V}\left[X|Y\right]] + \mathbb{V}_Y[\mu_{X|Y}] \tag{2.48}$$

To get some intuition for these formulas, consider a mixture of K univariate Gaussians. Let Y be the hidden indicator variable that specifies which mixture component we are using, and let

[2] This example is from `https://en.wikipedia.org/wiki/Law_of_total_expectation`, but with modified notation.

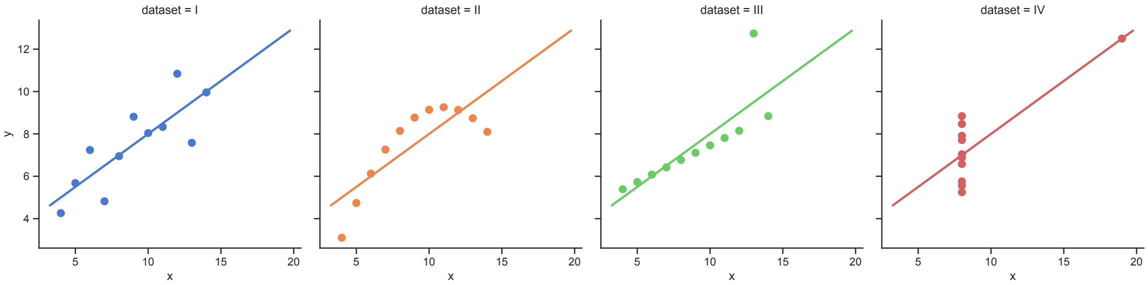


Figure 2.5: Illustration of Anscombe's quartet. All of these datasets have the same low order summary statistics. Generated by code at figures.probml.ai/book1/2.5.

$X = \sum_{y=1}^{K} \pi_y \mathcal{N}(X|\mu_y, \sigma_y)$. In Figure 2.4, we have $\pi_1 = \pi_2 = 0.5$, $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 0.5$. Thus

$$\mathbb{E}\left[\mathbb{V}\left[X|Y\right]\right] = \pi_1\sigma_1^2 + \pi_2\sigma_2^2 = 0.25 \tag{2.49}$$

$$\mathbb{V}\left[\mathbb{E}\left[X|Y\right]\right] = \pi_1(\mu_1 - \overline{\mu})^2 + \pi_2(\mu_2 - \overline{\mu})^2 = 0.5(0-1)^2 + 0.5(2-1)^2 = 0.5 + 0.5 = 1 \tag{2.50}$$

So we get the intuitive result that the variance of X is dominated by which centroid it is drawn from (i.e., difference in the means), rather than the local variance around each centroid.

2.2.6 Limitations of summary statistics *

Although it is common to summarize a probability distribution (or points sampled from a distribution) using simple statistics such as the mean and variance, this can lose a lot of information. A striking example of this is known as **Anscombe's quartet** [Ans73], which is illustrated in Figure 2.5. This shows 4 different datasets of (x, y) pairs, all of which have identical mean, variance and correlation coefficient ρ (defined in Section 3.1.2): $\mathbb{E}[x] = 9$, $\mathbb{V}[x] = 11$, $\mathbb{E}[y] = 7.50$, $\mathbb{V}[y] = 4.12$, and $\rho = 0.816$.[3] However, the joint distributions $p(x, y)$ from which these points were sampled are clearly very different. Anscombe invented these datasets, each consisting of 10 data points, to counter the impression among statisticians that numerical summaries are superior to data visualization [Ans73].

An even more striking example of this phenomenon is shown in Figure 2.6. This consists of a dataset that looks like a dinosaur[4], plus 11 other datasets, all of which have identical low order statistics. This collection of datasets is called the **Datasaurus Dozen** [MF17]. The exact values of the (x, y) points are available online.[5] They were computed using simulated annealing, a derivative free optimization method which we discuss in the sequel to this book, [Mur22]. (The objective function being optimized measures deviation from the target summary statistics of the original dinosaur, plus distance from a particular target shape.)

[3] The maximum likelihood estimate for the variance in Equation (4.51) differs from the unbiased estimate in Equation (4.38). For the former, we have $\mathbb{V}[x] = 10.00$, $\mathbb{V}[y] = 3.75$, for the latter, we have $\mathbb{V}[x] = 11.00$, $\mathbb{V}[y] = 4.12$.

[4] This dataset was created by Alberto Cairo, and is available at http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

[5] https://www.autodesk.com/research/publications/same-stats-different-graphs. There are actually 13 datasets in total, including the dinosaur. We omitted the "away" dataset for visual clarity.