1.7 Summarizing a distribution using moments and quantiles

To fully characterize the distribution of a random variable, we often write down objects like the cdf or the pmf/pdf, from which any probability can be computed. Having access to a complete description of the probability behavior of a random variable is nice, but we are often interested in extracting from the distribution of a random variable a few principle *features* or *summaries* of the distribution. We have already seen some:

- The expected value $\mu = \mathbb{E}[X]$ is a single number that averages over the distribution of X to answer "for all the random variation in X, where, at the end of the day, will its value typically be concentrated?"
- The variance $\sigma^2 = \text{var}(X) = \mathbb{E}[(X - \mu)^2]$ is a single number that answers "how far should I expect X to be from its expected value?"

When they exist, these are nice ways to concisely report on the key features of the random variation of X, without having to confront every detail of what the pmf/pdf is up to. In this lecture, we discuss a few more ways to summarize a distribution.

First, we can talk about the *shape* of the distribution; see Figure 1.27.

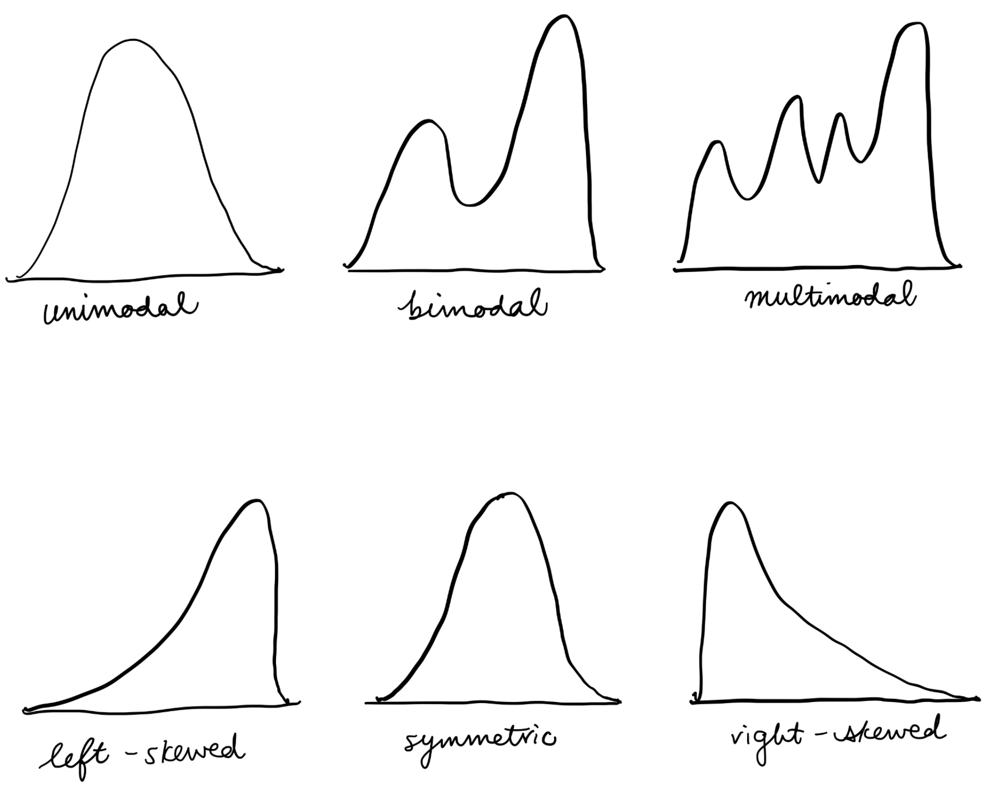

Figure 1.27: Shapes a distribution can take. Is there skew? Are there one or several modes? Note: left-skewed is also called "positively skewed" and right-skewed is also called "negatively skewed".
"

Example 1.19.
- People's heights form a bimodal distribution: women's heights are the left bump (to the left on the x-axis) and men's heights are the right bump (to the right on the x-axis).
- If someone doesn't show up for an exam and gets a zero, this can make the exam grade distribution left-skewed.
- Incomes in the U.S. are a right-skewed distribution (with Elon Musk and co. in the upper tail).

As you can see from the figure, the skew characterizes the degree of symmetry of a distribution. The below definition gives a quantitative measure of skew. We also introduce another way to characterize a distribution called the kurtosis.

Definition 1.20. For a random variable X with mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{var}(X) = \sigma^2$, the **skew** and **kurtosis** of X are defined as

$$\mathrm{skew}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \qquad \mathrm{kurtosis}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]. \tag{1.29}$$

Some remarks:

- While skew measures degree of asymmetry, kurtosis measure how heavy the tails of a distribution are. See Figure 1.28. Thicker tails mean extreme events are more likely.
- Skew and kurtosis are the third and fourth moments of the *standardized* version of X. We saw standardization in the last lecture in the context of Gaussian r.v.'s: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim \mathcal{N}(0,1)$. More generally, we can take any X with mean μ and variance σ^2, and standardize it to have mean 0 and variance 1 by taking $(X-\mu)/\sigma$.
- The fact that we have standardized X before taking moments means that skew and kurtosis do not care about location or scale. Shifting the distribution or spreading the points out does not affect skew or kurtosis.

1.7.1 Moments

We have seen lots of variations on "moments" so far. Here are the definitions for the different moments of a random variable X that has $\mathbb{E}[X] = \mu$ and $\mathrm{var}(X) = \sigma^2$:

n^{th} raw moment	$E(X^n)$
n^{th} absolute moment	$E(\lvert X\rvert^n)$
n^{th} central moment	$E[(X-\mu)^n]$
n^{th} absolute, central moment	$E(\lvert X-\mu\rvert^n)$
n^{th} standardized moment	$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^n\right]$
n^{th} absolute, standardized moment	$\mathbb{E}\left[\left\lvert\frac{X-\mu}{\sigma}\right\rvert^n\right]$

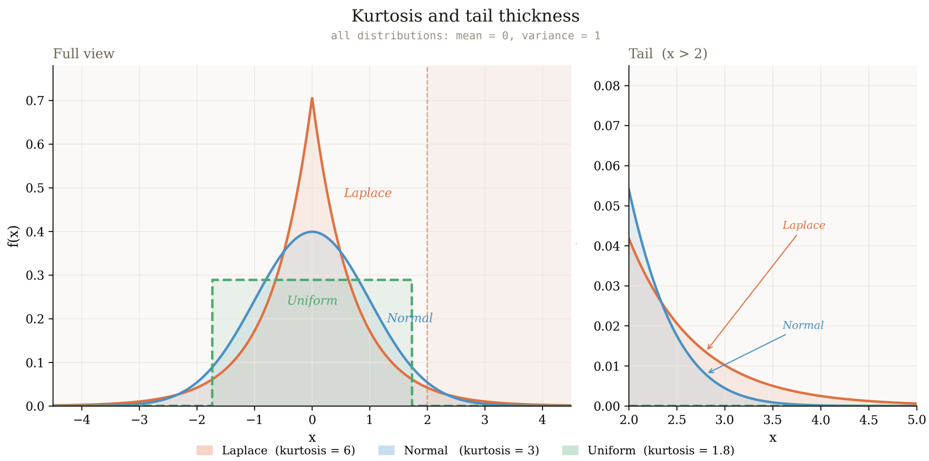

Figure 1.28: Higher kurtosis means heavier tails. The inset zooms in on the right tail, which is heavier for the Laplace distribution. The uniform distribution has the lightest tail, since the probability drops down to zero outside an interval.

So we see that the mean is the first raw moment, the variance is the second central moment, the skewness is the third centralized moment, and the kurtosis is the fourth standardized moment. This is summarized in Table 1.4.

Definition 1.21. We say a moment $\mathbb{E}[X^n]$ **exists** if the corresponding absolute moment is finite: $\mathbb{E}[|X|^n] < \infty$. So the mean exists, for example, if $\mathbb{E}[|X|] < \infty$.

1.7.2 Quantiles

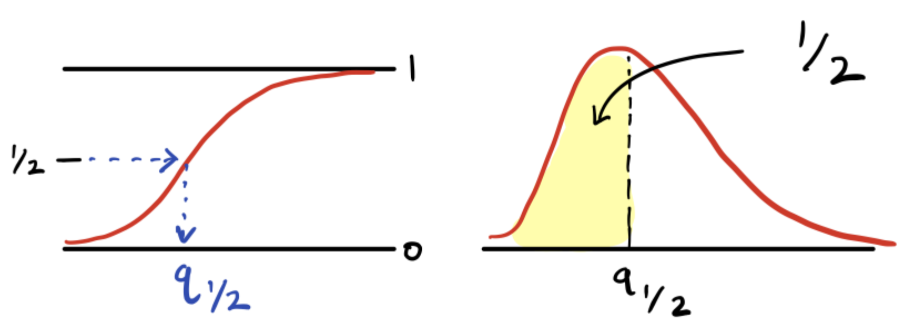

Figure 1.29

The mean and standard deviation have the significant drawback that they are not robust to outliers. Consider adding Elon Musk's income to the distribution of incomes in America. This will drastically increase both the mean and the standard deviation! *Quantiles*, on the other hand, are robust to outliers.

	raw	central	standardized
$n = 1$	$\mu = \mathbb{E}[X]$ "mean"	0	0
$n = 2$	$E(X^2)$	$\sigma^2 = E[(X - \mu)^2]$ "variance"	1
$n = 3$	$\mathbb{E}[X^3]$	$\mathbb{E}[(X - \mu)^3]$	$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ "skewness"
$n = 4$	$E(X^4)$	$E[(X - \mu)^4]$	$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ "kurtosis"
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1.4

Definition 1.22. An αth quantile is any real number q_α satisfying

$$P(X \leq q_\alpha) = \alpha.$$

In particular, the **median** is the quantile $q_{1/2}$.

So q_α is a point in the range with α probability to the left of it. Note that quantiles are not unique in general. When the cdf F_X is an invertible function (no jumps or plateaus), then the quantiles are unique, and you can define them as

$$q_\alpha = F_X^{-1}(\alpha).$$

For this reason, the inverse cdf F_X^{-1} is called the **quantile function**. See Figure 1.29 and 1.30.

Example 1.20. Consider the discrete distribution $X \sim \text{Unif}(1.2, 3.4, 5, 6.1, 8.9, 10)$, in which each of these numbers has probability $1/6$. Then e.g. $q_{5/6} = 8.9$, and the median is $q_{1/2} = 5$.

Consider replacing the 10 with 100. This will have a significant effect on the mean. Previously, $\mathbb{E}[X] = (1.2 + 3.4 + 5 + 6.1 + 8.9 + 10)/6 = 5.767$. With 100 instead of 10, we have $\mathbb{E}[X] = (1.2 + 3.4 + 5 + 6.1 + 8.9 + 100)/6 = 20.767$. But the median remains the same: it's $q_{1/2} = 5$ either way.

The above example shows that the median is a robust measure of location. What about a robust measure of spread? We can also use quantiles for that.

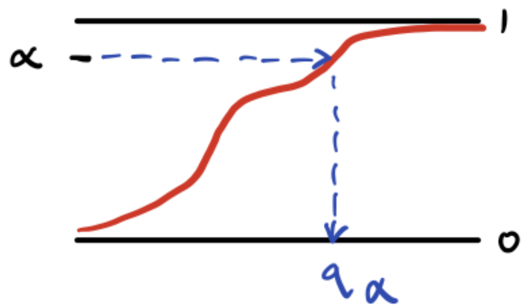

Figure 1.30

Definition 1.23. The interquartile range (IQR) is defined as

$$\mathrm{IQR} = q_{3/4} - q_{1/4}.$$

In words, the IQR is the width of the interval containing the middle 50% of the data. We summarize these different measures below.

	location	spread
not robust	mean	std. dev.
robust	median	IQR

Example 1.21. Let $X \sim \mathrm{Exp}(\beta)$. Recall from last lecture that the cdf of $\mathrm{Exp}(\beta)$ is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\beta x}, & x \geq 0. \end{cases}$$

Let us compute q_α for some $\alpha \in (0, 1)$. We solve

$$\begin{aligned} \alpha &= 1 - e^{-\beta q_\alpha} \\ e^{-\beta q_\alpha} &= 1 - \alpha \\ -\beta q_\alpha &= \ln(1 - \alpha) \\ q_\alpha &= -\frac{1}{\beta} \ln(1 - \alpha) \end{aligned}$$

and so the quantile function is

$$F_X^{-1}(\alpha) = -\frac{1}{\beta} \ln(1 - \alpha).$$

The median of the exponential distribution is

$$F_X^{-1}(0.5) = -\frac{\ln(1 - 0.5)}{\beta} = -\frac{\ln(0.5)}{\beta} = \frac{\ln 2}{\beta}.$$