1. **Order statistics**. Suppose there is a baseball player who is a free agent and thus able to sign with any team. Nine teams are interested. The player's actual value is 1/2, but none of the teams involved know this for sure; all of them can only estimate his value. Each team independently comes up with an estimate X_i, which are iid random variables uniformly distributed on $[0, 1]$. Let Y be the second highest of the X_i.

   (a) Convince yourself that Y is a good estimate for what the player actually gets paid. (Consider an auction between the nine teams. When does each team drop out?)

   (b) Compute $\mathbb{P}(Y \leq 2/3)$. (Hint: there are two qualitatively different ways that this can happen). Don't worry about simplifying your answer.

   (c) Find the CDF of Y.

   (d) Find the PDF of Y.

   (e) Compute $\mathbb{E}[Y]$. By how much is the player 'overpaid' on average? If each individual team was being rational, why does this occur?

   This phenomenon, where the winner of an auction tends to have overpaid, is known as the *winner's curse*.

2. **Sample variance**. We'll show why the sample variance has an $n-1$ in the denominator instead of an n. Let $X_1, \ldots, X_n$ be iid random variables with finite mean and variance. Define $\bar{X}$ to be the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$. Then, let $\tilde{s}^2 = \frac{1}{k}\sum_{i=1}^{n}(X_i - \bar{X})^2$, for k to be determined. We want to explain why $n-1$ is a better choice for k.

   (a) Show that $\mathbb{E}[\bar{X}] = \mathbb{E}[X_i]$ and that $\operatorname{Var}[\bar{X}] = \frac{1}{n}\operatorname{Var}[X_i]$.

   (b) Show that $\mathbb{E}[\tilde{s}^2] = \frac{n-1}{k}\operatorname{Var} X_i$. Linearity of expectation is the way to success, but note that X_i and $\bar{X}$ are not independent. In other words, if we want to use $\tilde{s}^2$ to estimate the variance of X_i, we want $k = n-1$ so that the expectation lines up. This property is called having an *unbiased estimator*, and it is frequently (though not always) desirable in statistics.

   (c) From part (b), we can see that the uncorrected sample variance underestimates the actual variance. Why does this happen?

(1)

(a)

If we think of the nine teams as being participants in an auction to obtain the player, then each team will keep bidding until the price is at their estimation of the player's value. Once the second highest valuing team drops out, the bidding stops, and therefore the player will get that much money.

(b)

There are two ways that the second highest estimation can be less than 2/3.

- There can be exactly one higher than 2/3. This has probability $9 \cdot (2/3)^8 \cdot (1/3)^1$; we pick the team that has the high estimate in one of 10 ways, and then the uniform distribution fills in the rest.
- All of them can be less than 2/3. This has probability $(2/3)^9$.

Since these are disjoint, the probability is just

$$9\left(\frac{2}{3}\right)^8 \cdot \frac{1}{3} + \left(\frac{2}{3}\right)^9 = \frac{2816}{19683} \approx 0.1431$$

(c)

We follow the same line of reasoning as in part (b):

$$F_Y(t) = \mathbb{P}(Y \le t) = 9t^8(1-t) + t^9 = 9t^8 - 8t^9 \qquad (0 \le t \le 1)$$

(d)

The PDF is just the derivative of the CDF:

$$f_y(t) = \frac{d}{dt}(9t^8 - 8t^9) = 72(t^7 - t^8) \qquad (0 \le t \le 1).$$

Note that for $t \in (0,1)$, we have $t^7 > t^8$, so this PDF is actually positive.

(e)

The expectation can be found by integration:

$$\begin{aligned}
\mathbb{E}[Y] &= \int t f_y(t)\ dt \\
&= \int_0^1 72(t^8 - t^9)\ dt \\
&= 72\left[\frac{t^9}{9} - \frac{t^{10}}{10}\right]_0^1 \\
&= 72(1/9 - 1/10) = \frac{72}{90} = 4/5 = 0.8
\end{aligned}$$

So the player will on average be overpaid by 0.3. Each individual team acted reasonably, but the team that won the auction is more likely to be one who overvalued the player.

(2)

(a)

Using the properties of expectation and variance:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_i]$$

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{\mathrm{Var}[X_i]}{n}$$

Note that we use independence of the X_i in the second computation.

(b)

We compute:

$$\begin{aligned}
\mathbb{E}[\tilde{s}^2] &= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{k}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - 2X_i\bar{X} + \bar{X}^2\right] \\
&= \frac{1}{k}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[X_i\bar{X}] + \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{k}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[X_i\bar{X}] + \mathrm{Var}[\bar{X}] + \mathbb{E}[\bar{X}]^2
\end{aligned}$$

Note that the expectation $\mathbb{E}[X_i\bar{X}]$ is not the same as $\mathbb{E}[X_i]\mathbb{E}[\bar{X}]$ since they're not independent. So we compute it separately:

$$\begin{aligned}
\mathbb{E}[X_i\bar{X}] &= \mathbb{E}\left[X_i \cdot \frac{1}{n}\sum_{j=1}^{n} X_j\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[X_iX_j\right] \\
&= \frac{1}{n}\left(\mathbb{E}[X_i^2] + \sum_{j\neq i}^{n}\mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]\right) \\
&= \frac{1}{n}\mathbb{E}[X_i^2] + \frac{n-1}{n}\mathbb{E}[X_i]^2
\end{aligned}$$

Plug this and the values from part (a) back in to the original expression to get

$$\begin{aligned}
\mathbb{E}[\tilde{s}^2] &= \frac{1}{k}\sum_{i=1}^{n} \mathbb{E}[X_i^2] - \frac{2}{n}\mathbb{E}[X_i^2] - \frac{2(n-1)}{n}\mathbb{E}[X_i]^2 + \frac{1}{n}\operatorname{Var}[X_i] + \mathbb{E}[X_i]^2 \\
&= \frac{1}{k}\sum_{i=1}^{n} \frac{n-2}{n}\mathbb{E}[X_i^2] - \frac{n-2}{n}\mathbb{E}[X_i]^2 + \frac{1}{n}\operatorname{Var}[X_i] \\
&= \frac{1}{k}\sum_{i=1}^{n} \frac{n-2}{n}\operatorname{Var}[X_i] + \frac{1}{n}\operatorname{Var}[X_i] \\
&= \frac{1}{k}\sum_{i=1}^{n} \frac{n-1}{n}\operatorname{Var}[X_i] \\
&= \frac{n-1}{k}\operatorname{Var}[X_i]
\end{aligned}$$

So if we want the expectation to be the same as $\operatorname{Var}[X_i]$, we have to let $k = n - 1$.

(c)

If we take a sample $X_1, \ldots, X_n$, then if we compute the dispersion around the *population* mean, this would be a good estimate of the population variance. But since we don't know the population mean, we are computing the dispersion around the sample mean. And the sample is clustered closer to the sample mean than the population mean. So that will probably be an underestimate.