170S Week 2 Discussion Notes

Colin Ni

January 15, 2025

Warm-up

Let v, w be two distinct points in $\mathbb{R}^m$. Find a constant-speed parametrization $\ell(t)$ of the line passing through v and w such that $\ell(0) = v$ and $\ell(1) = w$.

Solution. Set

$$\ell(t) = (1-t)v + tw.$$

It is easy to check that $\ell(0) = v$ and $\ell(1) = w$, and this has constant speed $\|\ell'(t)\| = \|(-1)v + w\|$. $\square$

Sample percentiles

Let $x_1, \ldots, x_n \in \mathbb{R}$ be a sample, and denote by $x_{(1)}, \ldots, x_{(n)}$ the order statistics of this sample.

Recall. The $(100p)$*th sample percentile* $\tilde{\pi}_p$ of this sample is not defined for every value $p \in [0, 1]$ but rather only for $p \in [\frac{1}{n+1}, \frac{n}{n+1}]$. To define $\tilde{\pi}_p$, write $(n+1)p$ uniquely as $r + t$ where $r \in \mathbb{Z}$ and $t \in [0, 1)$, and set

$$\tilde{\pi}_p = (1-t)x_{(r)} + tx_{(r+1)}.$$

Notice the similarity between $\tilde{\pi}_p$ and the answer to the warm-up. Let us spell out the intuition.

Intuition. The sample percentile $\tilde{\pi}_p$ is a constant-speed path from $x_{(r)}$ to $x_{(r+1)}$ on the interval $p \in [\frac{r}{n+1}, \frac{r+1}{n+1}]$. In particular $\tilde{\pi}_{\frac{r}{n+1}} = x_{(r)}$.

Example. Consider a sample such that

$$x_{(1)} = -2, \qquad x_{(2)} = 0, \qquad x_{(3)} = \frac{1}{2}, \qquad x_{(4)} = \frac{7}{2}.$$

Then $\tilde{\pi}_p$ is defined for $p \in [\frac{1}{5}, \frac{4}{5}]$, and the values $\tilde{\pi}_{\frac{1}{5}}$, $\tilde{\pi}_{\frac{2}{5}}$, $\tilde{\pi}_{\frac{3}{5}}$, $\tilde{\pi}_{\frac{4}{5}}$ are simply $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$. You should check that the 65th percentile $\tilde{\pi}_{0.65}$ is $\frac{5}{4}$, and you should draw a picture that shows that as p increases from $\frac{1}{5}$ to $\frac{4}{5}$, the speed of $\tilde{\pi}_p$ is $\frac{20}{2}$, then $\frac{5}{2}$, then $\frac{15}{2}$.

Sample mean and variance

Let again $x_1, \dots, x_n \in \mathbb{R}$ be a sample.

Recall. The *sample mean* and the *sample variance* of this sample are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

As a quick exercise, let us do a problem from the homework.

Exercise. Consider a linear transformation $y_i = ax_i + b$ of this sample, where $a, b \in \mathbb{R}$. Show that $\bar{y} = a\bar{x} + b$ and $s_y^2 = a^2 s_x^2$.

Solution. At once,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} b = a\bar{x} + b$$

and

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i + b - (a\bar{x} + b))^2 && \text{(using that } \bar{y} = a\bar{x} + b) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 \\
&= \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= a^2 s_x^2,
\end{aligned}$$

as required. □

One should wonder why we use $\frac{1}{n-1}$ instead of $\frac{1}{n}$ when computing the sample variance. This correction (multiplying the naive estimator by $\frac{n}{n-1}$) is called Bessel's correction, and it makes the estimator unbiased, as we will now explain.

Bessel's correction. Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. The estimators

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

of μ and σ^2 are unbiased.

Proof. Being unbiased means $\mathbb{E}[\bar{x}] = \mu$ and $\mathbb{E}[s_X^2] = \sigma^2$. For the former, observe that

$$\begin{aligned}
\mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] && \text{(linearity of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu && \text{(the } X_i \text{ are i.i.d. with mean } \mu\text{)} \\
&= \mu,
\end{aligned}$$

and in fact we can compute its variance as

$$\begin{aligned}
\mathrm{Var}(\bar{x}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) && \text{(basic properties of Var)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 && \text{(the } X_i \text{ are i.i.d. with variance } \sigma^2\text{)} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

For the latter, observe that the naive estimator can be written as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}((X_i - \mu) - (\bar{x} - \mu))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((X_i - \mu)^2 - 2(X_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2) && \text{(expand)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{x} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - \mu)^2 \\
& && \text{(distribute the sum)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{x} - \mu)^2 + (\bar{x} - \mu)^2 && \text{(definition of } \bar{x}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{x} - \mu)^2,
\end{aligned}$$

so its expected value is

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{x})^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{x}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2]-\mathbb{E}[(\bar{x}-\mu)^2] \\
&\qquad \text{(linearity of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i)-\mathrm{Var}(\bar{x}) \\
&\qquad \text{(the definition } \mathrm{Var}(Y)=\mathbb{E}[(Y-\mathbb{E}[Y])^2] \text{ of variance, and } \mathbb{E}[\bar{x}]=\mu) \\
&= \sigma^2-\frac{\sigma^2}{n} \\
&\qquad \text{(the } X_i \text{ are i.i.d. with variance } \sigma^2\text{, and } \mathrm{Var}(\bar{x})=\tfrac{\sigma^2}{n}) \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}
$$

Thus the correction of multiplying by $\frac{n}{n-1}$ makes the expected value σ^2. $\square$