Averages of Random Variables (Chapters 8&9)

(as long as they are identically and independently distributed), or, Why Statistics Work

Suppose that $X_1, X_2, X_3, \ldots$ of random variables which are independent and have exactly the same distribution. In particular, they all have the same expected value $E(X_i) = \mu$ and the same variance $V(X) = \sigma^2$.

The average of the first n of these is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$= \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n .$$

$$E(\overline{X}_n) = E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right)$$

$$= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n)$$

$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu$$

$$= \frac{n}{n}\mu = \mu$$

$$V(\bar{X}_n) = V\left(\tfrac{1}{n}X_1 + \tfrac{1}{n}X_2 + \cdots + \tfrac{1}{n}X_n\right)$$ since the X_i are independent

$$= V\left(\tfrac{1}{n}X_1\right) + V\left(\tfrac{1}{n}X_2\right) + \cdots + V\left(\tfrac{1}{n}X_n\right)$$

$$= \tfrac{1}{n^2}V(X_1) + \tfrac{1}{n^2}V(X_2) + \cdots + \tfrac{1}{n^2}V(X_n)$$

$$= \tfrac{1}{n^2}\sigma^2 + \tfrac{1}{n^2}\sigma^2 + \cdots + \tfrac{1}{n^2}\sigma^2$$

$$= \tfrac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Notice that as n increases, the variance of the average decreases, so we expect the average to be closer to the expected value of μ.

This idea is theoretically summarized in the Laws of Large Numbers.

The text only gives what others call the Weak Law of Large Numbers.

(Weak) Law of Large Numbers (Ch. 8)

Suppose that $X_1, X_2, X_3, \ldots$ is a sequence of independent and identically distributed random variables with $E(X_i) = \mu$. Then, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left| \overbrace{\frac{X_1 + \cdots + X_n}{n}}^{\overline{X}_n} - \mu \right| \geq \varepsilon \right) = 0 .$$

(Strong) Law of Large Numbers (Not in textbook.)

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of independent and identically distributed random variables with $E(X_i) = \mu$. Then

$$P\left(\lim_{n\to\infty} \overbrace{\frac{X_1 + \cdots + X_n}{n}}^{\overline{X}_n} = \mu \right) = 1 .$$

This is why statistics should work: the probability that the average of your data points approaches the true expected value as the number of data points increases is 1.

More computationally useful: relate things to the standard normal distribution.

We'll modify our random variable to have the same expected value and variance as the standard normal distribution.

If X is a random variable with expected value μ and variance σ^2, the <u>normalization of X</u> is $Z = \frac{X-\mu}{\sigma}$. Then

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu)$$
$$= \frac{1}{\sigma}\left(E(X)-E(\mu)\right) = \frac{1}{\sigma}\left(E(X)-E(X)\right) = 0$$

& $V\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(X-\mu) = \frac{1}{\sigma^2}V(X) + \frac{1}{\sigma^2}V(-\mu)$ (constants have variance 0)

$$= \frac{1}{\sigma^2}V(X) = \frac{1}{\sigma^2}\sigma^2 = 1$$

just like for the standard normal.

Suppose that $X_1, X_2, X_3, \ldots$ is a sequence of random variables that are independent and identically distributed, with $E(X_i) = \mu$ and variance $V(X_i) = \sigma^2$.

Let $S_n^* = \dfrac{Z_1 + Z_2 + \cdots + Z_n}{\sqrt{n}}$

$$= \frac{\frac{1}{\sigma}(X_1 - \mu) + \frac{1}{\sigma}(X_2 - \mu) + \cdots + \frac{1}{\sigma}(X_n - \mu)}{\sqrt{n}}$$

$$= \frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)}{\sigma\sqrt{n}}$$

$$= \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}}$$

$$E(S_n^*) = E\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right)$$

$$= \frac{1}{\sigma\sqrt{n}} E(X_1 + \cdots + X_n - n\mu)$$

$$= \frac{1}{\sigma\sqrt{n}} \left(E(X_1) + \cdots + E(X_n) - n\mu\right)$$

$$= \frac{1}{\sigma\sqrt{n}} \cdot 0 = 0$$

($\mu + \cdots + \mu = n\mu$)

$$V(S_n^*) = V\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right)$$

$$= \frac{1}{(\sigma\sqrt{n})^2} V(X_1 + \cdots + X_n - n\mu)$$

$$= \frac{1}{\sigma^2 n}\left[V(X_1) + \cdots + V(X_n) + \underbrace{V(-n\mu)}_{0}\right]$$ since constants have no variance...

$$= \frac{1}{\sigma^2 n} \cdot \left[\sigma^2 + \cdots + \sigma^2\right] = \frac{n\sigma^2}{\sigma^2 n} = 1$$

This brings us to the <u>Central Limit Theorem</u>

If $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, then

$$\lim_{n\to\infty} P(a \le S_n^* \le b) = \lim_{n\to\infty} P\left(a \le \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le b\right)$$

$$= \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$$

<u>i.e.</u> the distribution of S_n^* approaches that of the standard normal distribution as n increases.

<u>Q:</u> How big does n have to be for the standard normal to give useful approximations?

<u>A.:</u> Empirically, about $n = 30$, provided there isn't some systematic bias. More is better!

Example: We toss a fair coin 30 times. (X counts heads.)
Each toss has expected value $\frac{1}{2}$
& variance $\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$ $\quad \sigma = \sqrt{\frac{1}{4}} = \frac{1}{2}$

$$P(13 \le X_1 + \cdots + X_{30} \le 15) \qquad S_{30}^* = \frac{X_1 + \cdots + X_{30} - 30\cdot\frac{1}{2}}{\frac{1}{2}\sqrt{30}}$$

$$= P\left(\frac{13-15}{\frac{1}{2}\sqrt{30}} \le S_{30}^* \le \frac{15-15}{\frac{1}{2}\sqrt{30}}\right)$$

$$= P\left(\frac{-2}{\frac{1}{2}\sqrt{30}} \le S_{30}^* \le 0\right)$$

$$= P\left(\frac{-4}{\sqrt{30}} \le S_{30}^* \le 0\right)$$

$$\approx P\left(\frac{-4}{5.5} \le S_{30}^* \le 0\right) \approx P(-0.7303 \le S_{30}^* \le 0)$$

$\downarrow$ standard normal

$$\approx P(-0.7303 \le Z \le 0)$$

$$= P(Z \le 0) - P(Z < -0.7303)$$

$$\approx 0.5 - 0.2327 \approx 0.2673$$