Math 180A, Fall 2005, Prof. Tesler – October 26, 2005
Normal distribution and long-term averages or sums

1. Mean and standard deviation of sums or averages of i.i.d. variables

Let X be a random variable with $\mu = E(X)$ and $\sigma = \text{SD}(X)$. Let $X_1, \ldots, X_n$ be n i.i.d. (independent identically distributed) random variables with the distribution of X.

Let $S_n = X_1 + \cdots + X_n$ be their sum and $\overline{X}_n = (X_1 + \cdots + X_n)/n = S_n/n$ be their average. Then

Sums	Averages
$E(S_n) = E(X_1) + \cdots + E(X_n) = n\,E(X_1) = n\mu$	$E(\overline{X}_n) = \mu$
$\text{Var}(S_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = n\,\text{Var}(X_1) = n\sigma^2$	$\text{Var}(\overline{X}_n) = \sigma^2/n$
$\text{SD}(S_n) = \sigma\sqrt{n}$	$\text{SD}(\overline{X}_n) = \sigma/\sqrt{n}$

To reduce confusion between the different types of standard deviation, we sometimes say "the *standard error* (SE) of the sum is $\sigma\sqrt{n}$" and "the *standard error* (SE) of the average is $\sigma/\sqrt{n}$."

2. Limit theorems

When n is large enough, the sum or average of n i.i.d. random variables (for *any* distribution) closely resembles the normal curve! The details depend on the specific distribution. The proofs of the first 3 results below are in the book, and the 4th is beyond the scope of the book (but Chap. 2.3 has some details).

(1) **Markov's inequality (p. 174):** If $X \geq 0$ then $P(X \geq a) \leq E(X)/a$ for every $a > 0$.

(a) **Example:** The average of three nonnegative numbers is 10. What's the largest that any of them could be? Set two of them to 0 and the third one to 30.

(b) Redo (a) if a weighted average were used instead of a regular average. Then you can make two numbers be 0, with combined probability $1 - p$ (some $0 \leq p \leq 1$), and set the third number to $30/p$ with probability p. The third number can be as high as you want, at the expense of reducing its probability.

(c) Instead of just 3 numbers, consider all distributions with $X \geq 0$ and $E(X) = 30$. How large can $P(X \geq 200)$ be?

$$\begin{aligned} E(X) &= \sum_{0 \leq x < 200} xP(X = x) + \sum_{x \geq 200} xP(X = x) \\ &\geq \sum_{0 \leq x < 200} 0P(X = x) + \sum_{x \geq 200} 200P(X = x) \geq 200P(X \geq 200) \end{aligned}$$

so $P(X \geq 200) \leq E(X)/200 = 30/200 = 0.15$.

(2) **The Law of Large Numbers** (called "The Law of Averages" in our book, p. 195): Define $\overline{X}_n$ as above. For any $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(|\overline{X}_n - \mu| < \epsilon\right) = 1$$

Interpretation: For either example on the back, pick a narrow interval of real numbers centered around the mean. For example, for the die, pick $(3, 4)$ or $(3.4, 3.6)$ or $(3.49, 3.51)$ (corresponding to $\epsilon = 1, 0.1, 0.01$). As n increases, notice the probability of the average being in that interval increases towards 1.

(3) **Chebyshev's inequality (p. 191):** For any random variable X and any (real) $k > 0$,

$$P\left(|X - E(X)| \geq k\,\text{SD}(X)\right) \leq \frac{1}{k^2}$$

Or, in terms of the z-score $Z = (X - E(X))/\,\text{SD}(X) = (X - \mu)/\sigma$,

$$P(|Z| \geq k) \leq \frac{1}{k^2} \qquad \text{so, } P(|Z| \leq k) \geq 1 - \frac{1}{k^2}$$

So for *any* random variable, the probability of being between $\mu \pm \sigma$ is at least $1 - 1/1^2 = 0$; being within $\mu \pm 2\sigma$ is at least $1 - 1/2^2 = 0.75$; being within $\mu \pm 3\sigma$ is at least $1 - 1/3^2 \approx 0.88$; etc.

(4) **Central Limit Theorem (p. 196):** For n i.i.d. random variables with sum S_n and average $\overline{X}_n$ as defined above, and any real numbers $a < b$,

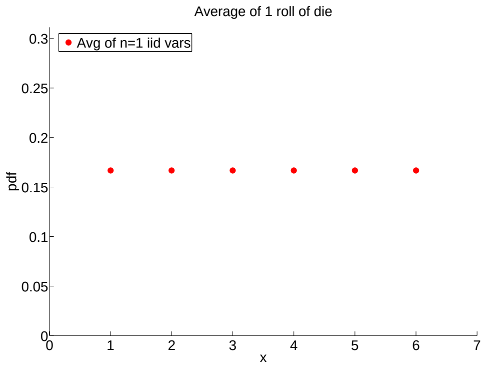
$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = P\left(a \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

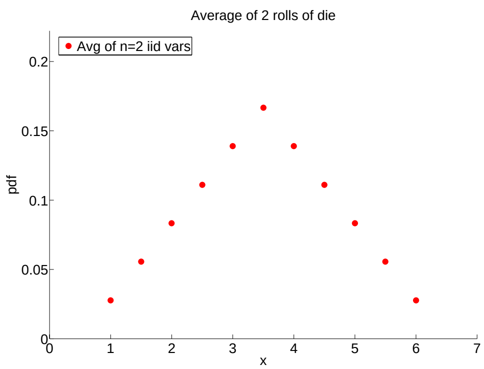
As $n \to \infty$, the approximation becomes exact equality.

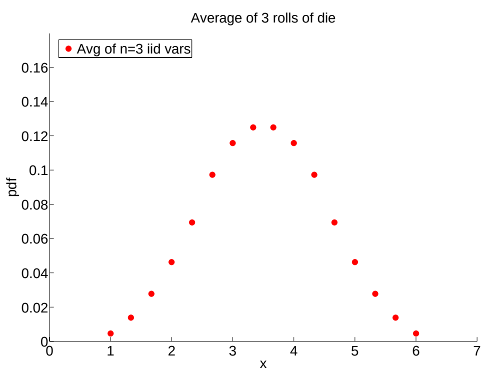
Interpretation: Notice in the two examples on the back, that as n increases, the pdf more and more closely resembles the normal curve. The mean of $\overline{X}_n$ stays the same as for X, while the standard deviation shrinks by the formula $\text{SD}(\overline{X}_n) = \text{SD}(X)/\sqrt{n}$.

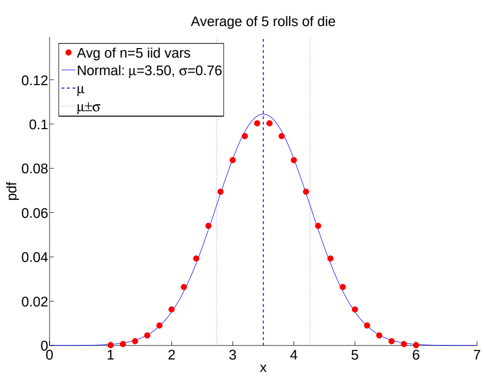
Repeated rolls of a die

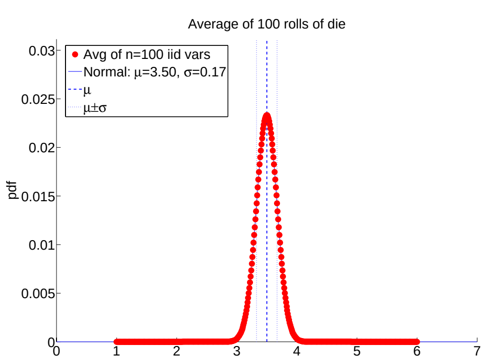
Initial distribution $\mu = 3.5$, $\sigma = \sqrt{35/12} \approx 1.71$











Repeated trials of “sawtooth” distribution

Initial distribution $\mu = 4$, $\sigma \approx 2.24$