3 Discrete random variables and distributions

In this section we introduce the concept of a *random variable*, a central notion in both probability and statistics. We will also study a number of important discrete probability distributions which are best described using random variables.

3.1 Discrete random variables

Definition 3.1 (Discrete random variable)**.** Let Ω be a discrete sample space. A function $X : \Omega \to \mathbb{R}$ is called a *discrete random variable* [1]. The image of X is denoted by S_X, i.e. $S_X := \{X(\omega) : \omega \in \Omega\} \subseteq \mathbb{R}$.

Given $A \subseteq \mathbb{R}$, the event $\{X \in A\} \subseteq \Omega$ is defined by

$$\big\{X \in A\big\} := \big\{\omega \in \Omega : X(\omega) \in A\big\}.$$

Given a probability distribution $\mathbb{P}$ on Ω we will write $\mathbb{P}(X \in A)$ for $\mathbb{P}(\{X \in A\})$.

Example 3.2 (Coins)**.** Suppose that we toss a coin n times. A natural sample space is $\Omega = \{H, T\}^n$. Then $X\big((\omega_1, \ldots, \omega_n)\big) = |\{1 \le i \le n : \omega_i = T\}|$ is a discrete random variable $X : \Omega \to \mathbb{R}$, which counts the number of tails appearing in the experiment. Here $S_X = \{0, \ldots, n\}$. If the coin is fair, then

$$\mathbb{P}\left(X = 0\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = 0\}\right) = \mathbb{P}\left(\{(H, H, \ldots, H)\}\right) = 2^{-n}.$$

Example 3.3 (Dice)**.** In many board games (e.g. Monopoly) a player first rolls two dice ⚀ ⚁ and then moves a number of steps equal to the sum of the two numbers on the dice. Let $\Omega = \{1, \ldots, 6\}^2$ be the sample space. Given $(\omega_1, \omega_2) \in \Omega$ the number of steps is given by $X\big((\omega_1, \omega_2)\big) = \omega_1 + \omega_2$. As X is a function from Ω to $\mathbb{R}$, it is a random variable.

Here $S_X = \{2, \ldots, 12\}$. If the dice are fair then $\mathbb{P}$ is the uniform distribution on Ω and

$$\mathbb{P}\big(X = x\big) = \mathbb{P}\big(\big\{(\omega_1, \omega_2) \in \Omega : X\big((\omega_1, \omega_2)\big) = x\big\}\big) = \begin{cases} \frac{x-1}{36} & \text{if } 2 \le x \le 6, \\ \frac{13-x}{36} & \text{if } 7 \le x \le 12. \end{cases}$$

Example 3.4 (Lottery)**.** An urn contains 49 balls, numbered 1 to 49. We draw 6 balls from the urn one at a time, replacing at each step. What is the probability the largest drawn number is at most k?

We take $\Omega = \{1, \ldots, 49\}^6$. Defining $X((\omega_1, \ldots, \omega_6)) = \max\{\omega_1, \ldots, \omega_6\}$, we are interested in $\mathbb{P}\left(X \le k\right)$. Note that $\{X \le k\} = \{(\omega_1, \ldots, \omega_6) \in \Omega : \omega_i \le k \text{ for all } i\}$. This set has k^6 elements. Hence, for $k = 0, 1, \ldots, 49$, we have

$$\mathbb{P}\left(X \le k\right) = \left(\frac{k}{49}\right)^6.$$

As the event $\{X = k\}$ is equal to $\{X \le k\} \setminus \{X \le k-1\}$, for all $k \in S_X = \{1, \ldots, 49\}$ we have

$$\mathbb{P}\left(X = k\right) = \mathbb{P}\left(\{X \le k\} \setminus \{X \le k-1\}\right) = \mathbb{P}\left(X \le k\right) - \mathbb{P}\left(X \le k-1\right) = \frac{k^6 - (k-1)^6}{49^6}.$$

[1] The term 'random variable' is very well established, but a bit unfortunate ... they are just functions.

Example 3.5 (Birthday paradox)**.** Recall the birthday paradox (Example 1.24). We had $\Omega = \{1, \ldots, 365\}^k$, where $\omega = (\omega_1, \ldots, \omega_k)$ records the birthdays of k people. Let $X(\omega) = |\{\omega_1, \ldots, \omega_k\}|$ be the number of different birthdays in the group. The event A in Example 1.24 is exactly the event $\{X = k\}$, i.e. there are no birthday collisions.

Remark 3.6. Although we have seen some examples above, random variables often seem a little mysterious when met for the first time. Here are some of the reasons why we study them.

(a) We are often more interested in a quantity or statistic associated to the outcome of a random experiment rather than the outcome itself, e.g. Example 3.3.

(b) We can use algebra to study them. For example given two random variables $X, Y : \Omega \to \mathbb{R}$ then $X+Y$, which maps $\omega \mapsto X(\omega)+Y(\omega)$, is also a random variable. Similarly we can consider $X \cdot Y$, $10X - 3Y$, ..., or take a function $f : \mathbb{R} \to \mathbb{R}$ and consider the random variable $f \circ X : \Omega \to \mathbb{R}$, given by $\omega \mapsto f(X(\omega))$, e.g. X^2 is a random variable. This will be important in later sections.

(c) Random variables give a simple way to describe many interesting probability distributions, as shown by the next result.

3.1.1 Mass and distribution functions

Lemma 3.7. Let $\mathbb{P}$ be a probability distribution on a discrete sample space Ω and let $X : \Omega \to \mathbb{R}$ be a discrete random variable with image S_X. Then the function $p_X : S_X \to [0, 1]$ defined by $p_X(x) := \mathbb{P}(X = x)$ is a discrete probability distribution on S_X.

Proof. As Ω is a discrete set, the same is also true of the image of Ω under X, i.e. S_X is discrete. We now use Lemma 1.8(b) to show p_X is a probability distribution on S_X. To see this, note that $p_X(x) = \mathbb{P}(\{X = x\}) \geq 0$ for all $x \in S_X$. Secondly, as $\Omega = \bigcup_{x \in S_X} \{X = x\}$ and these events are pairwise disjoint, we have $\sum_{x \in S_X} p_X(x) = \sum_{x \in S_X} \mathbb{P}(X = x) = \mathbb{P}(\Omega) = 1$ by Definition 1.6 (iii). Thus the statement follows by Lemma 1.8(b). □

Definition 3.8 (Mass and distribution function)**.** Let $X : \Omega \to \mathbb{R}$ be a discrete random variable. The *(probability) mass function* of X is the map $p_X : S_X \to [0, 1]$ given by

$$p_X(x) := \mathbb{P}(X = x).$$

The *(cumulative) distribution function* of X is the map $F_X : \mathbb{R} \to [0, 1]$ given for all $t \in \mathbb{R}$ by

$$F_X(t) := \mathbb{P}(X \leq t) = \sum_{x \in S_X, x \leq t} \mathbb{P}(X = x).$$

Example 3.9 (Dice)**.** For the random variable X from Example 3.3, we have $S_X = \{2, \ldots, 12\}$ and the mass function is given by

$$p_X(x) = \begin{cases} (x-1)/36 & \text{if } 2 \leq x \leq 6, \\ (13-x)/36 & \text{if } 7 \leq x \leq 12. \end{cases}$$

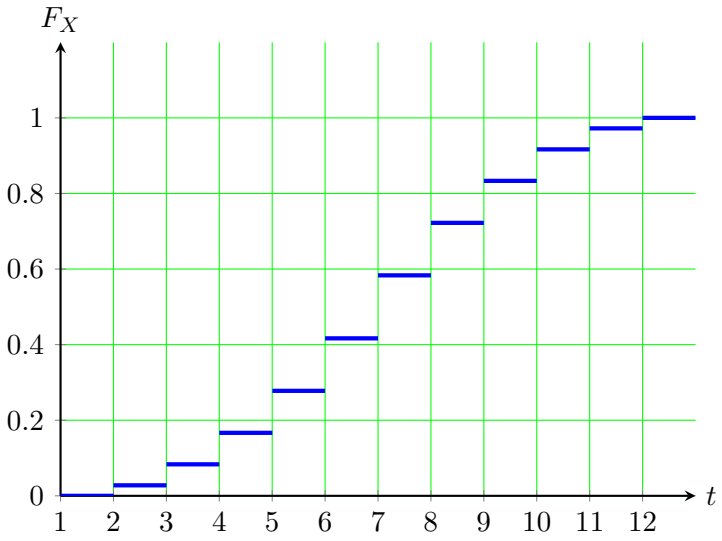

Figure 1: The distribution function of X from Example 3.9 drawn on the interval $[1, 13]$.

Example 3.10 (Discrete uniform distribution)**.** We say that a random variable X follows the uniform distribution on $\{1, \ldots, n\}$ if $S_X = \{1, \ldots, n\}$ and $\mathbb{P}(X = k) = 1/n$ for all $k \in \{1, \ldots, n\}$. Then [2]

$$F_X(t) = \begin{cases} 0 & \text{if } t < 1, \\ \frac{\lfloor t \rfloor}{n} & \text{if } 1 \le t \le n, \\ 1 & \text{if } t > n. \end{cases}$$

Example 3.11 (Lottery)**.** Reconsider Example 3.4. For the random variable X with $S_X = \{1, \ldots, 49\}$, we computed the mass function $p_X(x) = (x^6 - (x-1)^6)/49^6$ for $x \in S_X$. Its distribution function is

$$F_X(t) = \begin{cases} 0 & \text{if } t < 1, \\ \left(\frac{\lfloor t \rfloor}{49}\right)^6 & \text{if } 1 \le t \le 49, \\ 1 & \text{if } t > 49. \end{cases}$$

Example 3.12. Suppose that X is a discrete random variable with $S_X = \{-1, 4, 6\}$ and $p_X(-1) = 0.3, p_X(4) = 0.5$. What is F_X? First, note that $p_X(6) = 1 - p_X(-1) - p_X(4) = 0.2$. Then,

$$F_X(t) = \sum_{x \in S_X, x \le t} p_X(x) = \begin{cases} 0 & \text{if } t < -1, \\ 0.3 & \text{if } -1 \le t < 4, \\ 0.8 & \text{if } 4 \le t < 6, \\ 1 & \text{if } t \ge 6. \end{cases}$$

[2] $\lfloor x \rfloor$ is the largest integer not exceeding x (that is, x rounded down).

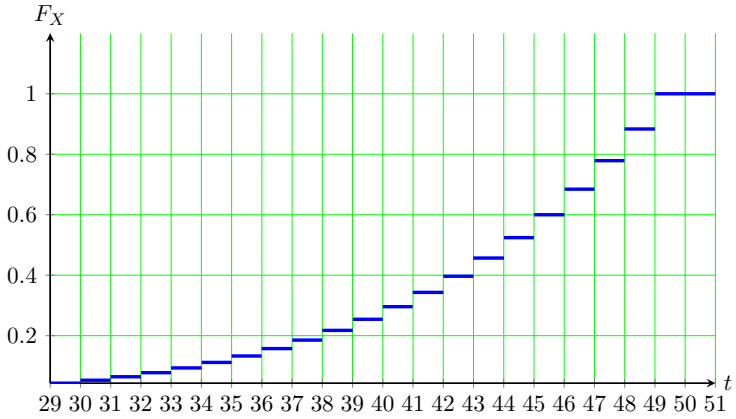

Figure 2: The distribution function of the random variable X from Example 3.4 and 3.11 drawn on $[29, 51]$.

Remark 3.13. When $S_X \subseteq \mathbb{Z}$ there is an easy way to obtain the mass function from its distribution function: for $k \in S_X$ the set $\{X \leq k\}$ is the disjoint union of $\{X = k\}$ and $\{X \leq k-1\}$ and so

$$p_X(k) = \mathbb{P}(\{X = k\}) = \mathbb{P}(\{X \leq k\}) - \mathbb{P}(\{X \leq k-1\}) = F_X(k) - F_X(k-1).$$

Similarly, if $S_X \subseteq \mathbb{Z}$ then $\{X \geq k\}$ is the disjoint union of $\{X = k\}$ and $\{X \geq k+1\}$, so

$$p_X(k) = \mathbb{P}(X = k) = \mathbb{P}(X \geq k) - \mathbb{P}(X \geq k+1).$$

As shown in the figures in this section, the distribution function $F_X(t)$ of a discrete random variable

- is constant between any two consecutive points $s < t$ in the set S_X.
- is continuous at $t \in \mathbb{R}$ if $t \notin S_X$.
- has an upward jump of magnitude $\mathbb{P}(X = t)$ at $t \in S_X$.

Proposition 3.14. Let X be a discrete random variable. For its distribution function F_X, we have

(i) F_X is monotonically increasing,

(ii) $\lim_{t \to -\infty} F_X(t) = 0$, and

(iii) $\lim_{t \to \infty} F_X(t) = 1$.

Proof. Monotonicity of F_X follows from Proposition 1.8(iii) since $\{X \leq s\} \subseteq \{X \leq t\}$ for $s \leq t$. Proofs of (ii) and (iii) are given in the module 2S in Year 2. □

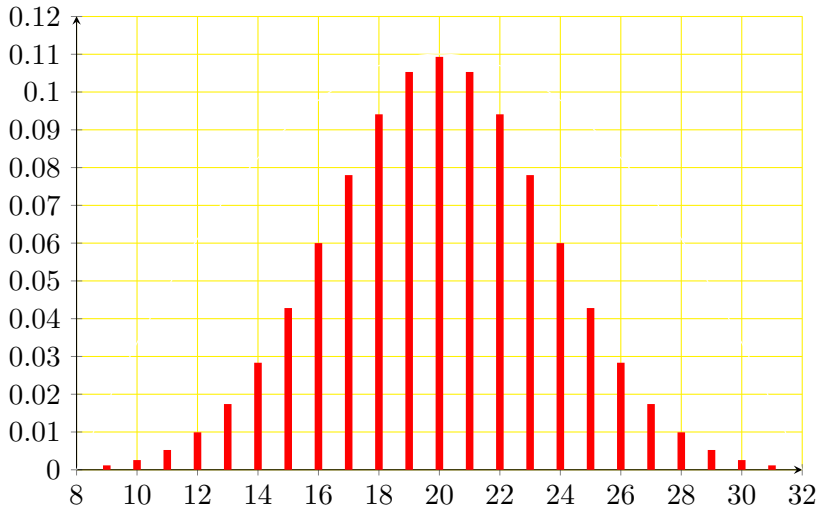

Figure 3: Mass function of $X \sim \text{bin}_{n,p}$ for $n = 60$ and $p = 1/3$ for $k = 9, \ldots, 31$.

3.2 The binomial distribution

The binomial distribution is the probability distribution which represents the number of successes in a repeated random experiment (e.g. n coin flips), where each attempt is successful with probability $p \in [0,1]$ and successes are independent. Sticking to coin flips, set $\Omega = \{H,T\}^n$ and choose the probability distribution $\mathbb{P}$ satisfying [3]

$$\mathbb{P}(\omega) = p^{|\{1\le i\le n:\omega_i=T\}|} \cdot (1-p)^{|\{1\le i\le n:\omega_i=H\}|} \quad \text{for } \omega = (\omega_1, \ldots, \omega_n) \in \Omega.$$

Let $X(\omega) = X\big((\omega_1, \ldots, \omega_n)\big) = |\{1 \le i \le n : \omega_i = T\}|$ be the number of tails. Then,

$$\mathbb{P}(X = k) = \mathbb{P}(\{\omega \in \Omega : |\{1 \le i \le n : \omega_i = T\}| = k\}) = \binom{n}{k} p^k (1-p)^{n-k},$$

since there are $\binom{n}{k}$ ways to choose the k positions $1 \le i \le n$ with $\omega_i = T$ [4].

Definition 3.15 (Binomial distribution)**.** The *binomial distribution with parameters* n and p, where $n \in \mathbb{N}$ and $p \in [0,1]$, is the probability distribution on $\{0, 1, \ldots, n\}$ given by

$$\text{bin}_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k \in \{0, \ldots, n\}.$$

A random variable X is said to follow the binomial distribution with parameters n and p if $\mathbb{P}(X = k) = \text{bin}_{n,p}(k)$ for $k \in S_X = \{0, 1, \ldots, n\}$. We denote this by writing $X \sim \text{bin}_{n,p}$.

By Proposition 1.8 this is a probability distribution on $\{0, \ldots, n\}$ as $\text{bin}_{n,p}(k) \ge 0$ for all $k \in \{0, \ldots, n\}$ and by the binomial theorem we have

$$\sum_{k=0}^{n} \text{bin}_{n,p}(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

[3] Note that, as the successes are independent, this probability here agrees with the calculation in Theorem 2.22.

[4] Indeed, this is an unordered sample, without repetition.

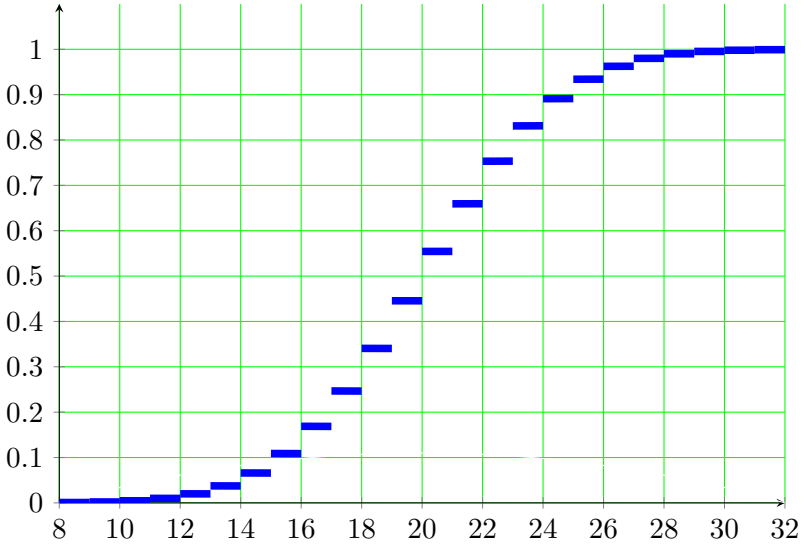

Figure 4: Distribution function of $X \sim \text{bin}_{n,p}$ for $n = 60$ and $p = 1/3$ in the range $8 \leq t \leq 32$.

Example 3.16 (Telephone calls)**.** When making a phone call in the old days the probability to establish a connection was 0.7. What was the probability that at least 5 of 10 calls were successful? Let X be the number of successful calls. Then $X \sim \text{bin}_{10,0.7}$. Therefore,

$$\mathbb{P}(X \geq 5) = \sum_{k=5}^{10} \text{bin}_{10,0.7}(k) = \sum_{k=5}^{10} \binom{10}{k} 0.7^k 0.3^{10-k} = 0.952\ldots$$

Example 3.17 (Puppies)**.** In a litter any puppy is equally likely to be male or female, independently of other puppies. What is the probability of four male and four female puppies in a litter of eight? Let X be the number of female puppies. Then, $X \sim \text{bin}_{8,0.5}$. Hence,

$$\mathbb{P}(X = 4) = 2^{-8}\binom{8}{4} = \frac{35}{128} = 0.273\ldots$$

3.3 The hypergeometric distribution

An urn contains $t = r + s$ balls, of which $r \geq 1$ red and $s \geq 0$ are blue. We select a sample of size n and are interested in the number of red balls in the sample. If the sample is chosen with replacement then this number follows a binomial distribution with parameters n and $p = r/t$. We instead consider selecting without replacement.

First note, without replacement, the size of the sample is at most the number of balls in the urn, i.e. $n \leq t$. It is helpful to label the red balls from 1 to r and the blue balls from $r + 1$ to t. Let Ω be the set of all subsets of $\{1, \ldots, t\}$ of size n. Here, we represent a subset $\omega \in \Omega$ of size n as tuple $(\omega_1, \ldots, \omega_n)$ with $1 \leq \omega_1 < \ldots < \omega_n \leq t$, where $\omega_1, \ldots, \omega_n$ are the numbers in ω. The random variable $X(\omega) = |\{1 \leq k \leq n : \omega_k \leq r\}|$ counts the number of red balls in the sample. As all outcomes $\omega \in \Omega$

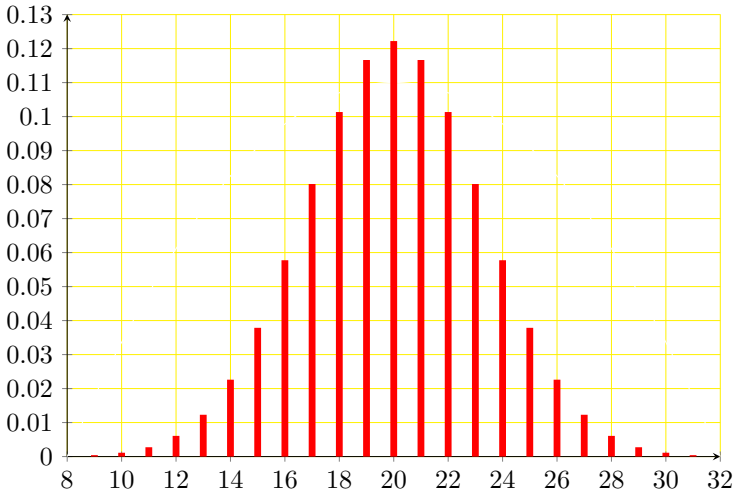

Figure 5: Mass function of $X \sim \mathrm{hyp}_{n,r,t}$ with $n = 60, r = 100$ and $t = 300$.

are equally likely, we consider the uniform distribution on Ω. Since there are exactly

$$\binom{r}{k}\binom{t-r}{n-k}$$

ways to choose a sample of size t with k balls with labels at most r (i.e. red balls), we obtain

$$\mathbb{P}(X = k) = \frac{\binom{r}{k}\binom{t-r}{n-k}}{\binom{t}{n}}.$$

Definition 3.18. The *hypergeometric distribution with parameters* $n, r, t \geq 1$ with $n, r \leq t$ is the probability distribution on $\{0, 1, \ldots, n\}$ given by

$$\mathrm{hyp}_{n,r,t}(k) = \frac{\binom{r}{k}\binom{t-r}{n-k}}{\binom{t}{n}}, \qquad \text{for } k \in \{0, 1, \ldots, n\}.$$

A random variable X is said to follow the hypergeometric distribution with parameters n, r, t if $\mathbb{P}(X = k) = \mathrm{hyp}_{n,r,t}(k)$ for $k \in S_X = \{0, 1, \ldots, n\}$. In this case we will write $X \sim \mathrm{hyp}_{n,r,t}$.

Since the values $\mathbb{P}(X = k)$ for $k \in \{0, 1, \ldots, n\}$ add up to one [5], our arguments show the useful formula

$$\sum_{k=0}^{n} \binom{r}{k}\binom{t-r}{n-k} = \binom{t}{n}.$$

[5] A pedantic point: the random variable X constructed at the beginning of this subsection in fact only has $S_X \subseteq \{0, 1, \ldots, n\}$. Indeed, equality cannot hold if n exceeds the number of blue balls (as then $X \geq 1$) or if n exceeds the number of red balls (then $X \leq r$). Formally, $S_X = \{i \geq 0 : \max(n - t + r, 0) \leq i \leq \max(n, r)\}$. However, distinguishing between S_X and $\{0, 1, \ldots, n\}$ is of no real significance as $\mathbb{P}(X = k) = 0$ for $k \in \{0, 1, \ldots, n\} \setminus S_X$.

Example 3.19 (Bridge)**.** In bridge, a 52 card deck is split among four players, so that each player receives 13 cards. Assuming that the deck is perfectly shuffled, what is the probability that the dealer obtains all four aces? The number of aces obtained by the dealer X follows the hypergeometric distribution with $t = 52, r = 4$ and $n = 13$. Hence,

$$\mathbb{P}(X = 4) = \frac{\binom{48}{9}}{\binom{52}{13}} = \frac{11}{4165} = 0.00264\ldots$$

Example 3.20 (Flashlight)**.** We find an old flashlight containing four batteries, all of which are dead. In attempting to replace them with four new batteries, we mix up old and new batteries so that we can not tell them apart. In frustration, we pick four and insert them. If the flashlight works provided at least three batteries are new, what is the probability that it works?

The number of new batteries X in our sample follows the hypergeometric distribution with $t = 8, r = 4$ and $n = 4$. The sought probability is

$$\mathbb{P}(X \geq 3) = \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{17}{70}.$$

3.4 The geometric distribution

The geometric distribution is the probability distribution which represents the time taken until the first success in a repeated random experiment, where in each iteration the probability of success is $p \in (0, 1)$ and all successes are independent.

A prototype example comes from flipping a coin repeatedly until the first tail appears, where tails appears in each flip with probability $p \in (0, 1)$. We could represent this with a sample space $\Omega = \{1, 2, \ldots\}$, where outcome $k \in \Omega$ means the first tail appears on the kth toss. To obtain outcome k requires head to appear in the first $k - 1$ tosses followed by a final tail. As each toss is independent, it is natural to set

$$\mathbb{P}(k) = p(1 - p)^{k-1}, \qquad \text{for } k \in \{1, 2, ...\}.$$

Definition 3.21 (Geometric distribution)**.** The *geometric distribution* [6] *with parameter* p, with $p \in (0, 1)$, is the probability distribution on $\{1, 2, \ldots\}$ given by

$$\text{geo}_p(k) = p(1 - p)^{k-1}, \qquad \text{for } k \in \{1, 2, \ldots\}.$$

A random variable X follows this distribution if $S_X = \{1, 2, \ldots\}$ and $\mathbb{P}(X = k) = \text{geo}_p(k)$ for $k \in S_X = \{1, 2, \ldots\}$. We then write $X \sim \text{geo}_p$.

These weights are non-negative, and $\sum_{k=1}^{\infty} p(1 - p)^{k-1} = 1$ using the geometric series formula (e.g. Example 1.20). As $\{1, 2, ...\} = \mathbb{N}$ is discrete, we have a probability distribution by Lemma 1.8.

[6] Many textbooks define the geometric distribution with mass function $p(1 - p)^k$ for $k = 0, 1, \ldots$. You could view this as the number of failures before the first success. As the number of failures differs by exactly one from the time of the first success, there is no conceptual difference between the two models – although this shift is a nuisance.

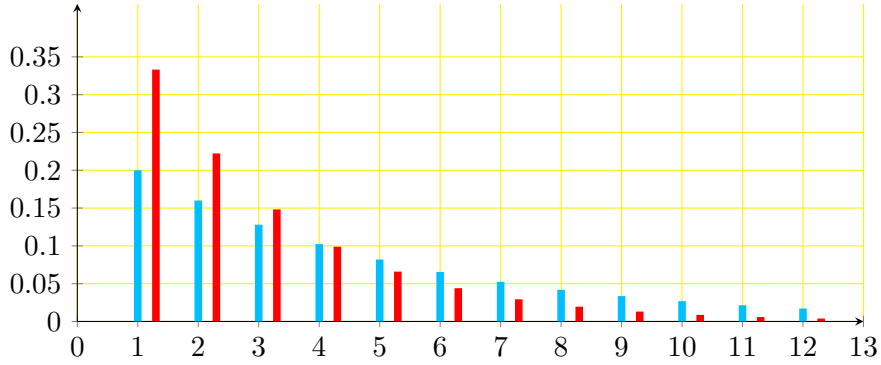

Figure 6: Mass function for $X \sim \text{geo}_p$ for $p = 1/5$ (left) and $p = 1/3$ (right).

The distribution function F_X is then given, for $t \in \mathbb{R}$, by

$$F_X(t) = \begin{cases} 1 - (1-p)^{\lfloor t \rfloor} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0. \end{cases}$$

Example 3.22 (Lottery)**.** The probability to win the National Lottery jackpot in the UK [7] is roughly $p = 1/\binom{49}{6} \approx 7.15 \cdot 10^{-8}$. How many times should you play to have at least 1% chance to win the jackpot at least once?

Let X denote the first time we would win the jackpot if we were to repeatedly play. Then $X \sim \text{geo}_p$. Therefore, if we wish to find $k \in \{1, 2, \ldots\}$ with $F_X(k) \geq 0.01$ we are lead to the calculation

$$0.99 \geq 1 - F_X(k) = (1-p)^k \quad \Longleftrightarrow \quad k \geq \frac{\log 0.99}{\log(1-p)} \approx 140564.$$

For context, if we were to play twice a week then we would need to play for more than 1350 years!

Example 3.23 (Birth control)**.** A couple decides to continue to have children until the first girl is born (and at that point, they stop). Let X denote the number of their children. Assuming that there are no medical reasons for which they would have to adjust their strategy, we have $X \sim \text{geo}_{0.5}$. Thus, the probability they have at least 3 boys is $\mathbb{P}(X \geq 4) = 1 - \mathbb{P}(X \leq 3) = 1/8$.

Proposition 3.24. Let $X \sim \text{geo}_p$ with $p \in (0, 1)$. Then

$$\mathbb{P}\big(X \geq n + m | X \geq n\big) = \mathbb{P}(X \geq m + 1) \qquad \text{for all } n \geq 1, m \geq 0. \tag{1}$$

Proof. For all $n \in \{1, 2, \ldots\}$ we have $\mathbb{P}(X \geq n) = 1 - F_X(n-1) = (1-p)^{n-1}$, and so

$$\mathbb{P}(X \geq n + m | X \geq n) = \frac{\mathbb{P}(X \geq n + m)}{\mathbb{P}(X \geq n)} = \frac{(1-p)^{n+m-1}}{(1-p)^{n-1}} = (1-p)^m = \mathbb{P}(X \geq m + 1),$$

as required. □

[7] This refers to the old 6 out of 49 game played until October 2015.

Remark 3.25. Property (1) is often referred to as the *memoryless* property of geometric distributions. Identity (1) becomes a little nicer if we consider $Y = X - 1$, that is, the number of failures before the first success. Then, the memoryless property states that

$$\mathbb{P}\big(Y \geq n + m | Y \geq n\big) = \mathbb{P}\left(Y \geq m\right), \quad \text{for all } n, m \in \{0, 1, 2, \ldots\}.$$

3.5 The Poisson distribution

Many natural quantities which arise in practice follow the Poisson distribution. At first it looks a little mysterious, but it can be seen as as an approximation for binomial distribution with parameters n and p if $p \approx \lambda/n$ for some $\lambda > 0$ and n is large.

Theorem 3.26 (Law of small numbers)**.** *Let $p_n, n \in \mathbb{N}$ be a sequence of real numbers with $0 \leq p_n \leq 1$ for all $n \in \mathbb{N}$ and $n \cdot p_n \to \lambda > 0$ as $n \to \infty$. Then, for all $k \in \{0, 1, \ldots\}$,*

$$\lim_{n\to\infty} \mathrm{bin}_{n,p_n}(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

A (non-examinable) sketch proof of Theorem 3.26 can be found in the Appendix at the end of this section.

Definition 3.27 (Poisson distribution [8])**.** Let $\lambda > 0$. The *Poisson distribution with parameter $\lambda > 0$* is the probability distribution on $\{0, 1, 2, \ldots\}$ given by

$$\mathrm{Poi}_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

A random variable X is said to follow the Poisson distribution with parameter λ if $\mathbb{P}\left(X = k\right) = \mathrm{Poi}_\lambda(k)$ for $k \in S_X = \{0, 1, 2, \ldots\}$. In this case we will write $X \sim \mathrm{Poi}_\lambda$.

The values associated with the Poisson distribution sum to one, using [9] that $e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$ so again it is a probability distribution by Lemma 1.8.

Example 3.28 (Buns with raisins)**.** A baker is preparing to bake 1000 buns with raisins. They add m raisins to the dough and knead the dough well so that the number of raisins in a bun follows a binomial distribution with parameters m and $1/1000$. Hence, if $m \sim 1000\alpha$ with $\alpha > 0$, the Poisson approximation can be used to describe the number of raisins in a bun. How many raisins should the baker add so that with probability at least 0.95 a chosen bun contains a raisin? We find

$$0.05 \geq \mathrm{Poi}_\lambda(0) = e^{-\alpha} \Leftrightarrow \alpha \geq \log 20 = 2.9957\ldots$$

Hence, they should add at least 3000 raisins.

[8] The Poisson distribution was first studied in 1837 by the French physicist and mathematician Siméon-Denis Poisson, but its importance was overlooked in Europe for a long time.

[9] This formula is **fundamental**. Keep it in mind.

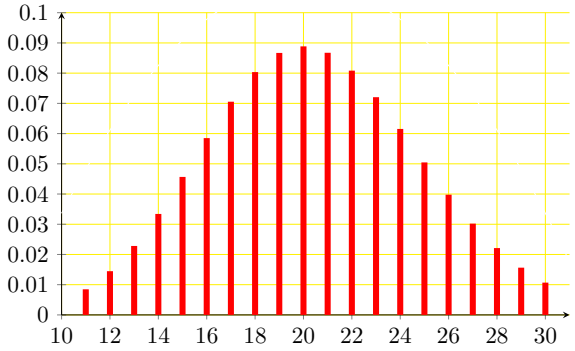

Figure 7: Mass function of $X \sim \text{Poi}_\lambda$ with $\lambda = 20.5$ for $k = 11, \ldots, 30$.

As $np_n \to \lambda$, the law of small numbers can be used even when n and p_n are unknown since only their product is relevant. This gives the theoretical justification why the Poisson distribution is used to model the occurrences of rare events.

The value λ then corresponds to the average value of the quantity under investigation [10].

Some examples of quantities which are well described by a Poisson distribution include:

- the annual number of fatal shark attacks worldwide (estimates give $4 \leq \lambda \leq 15$) [11];
- the annual number of observed gamma ray bursts;
- the number of shooting stars observed over a given time span in one night;
- the monthly number of major (≥ 7.0 on the Richter scale) earthquakes worldwide ($\lambda \approx 15$);
- the number of plants in parts of a forest; and
- the number of goals scored by a football team or a player in one match. [12]

Example 3.29 (Asteroids)**.** Asteroids of diameter about 1km hit the earth roughly every 500,000 years. We model the number of such strikes in the next 2 Million years by a Poisson random variable X with $\lambda = 4$ as the average value of such strikes in 2 Million years is four. The probability that no such strike occurs is $\mathbb{P}(X = 0) = e^{-4} = 0.0138\ldots$

Example 3.30 (Elevator failures)**.** An elevator is known to stop working once in four months on average. What are the chances that the elevator has at least three failures in the course of a whole

[10] Formally, the *expected value* of $X \sim \text{Poi}_\lambda$ is λ. We will make this connection in Section 5.

[11] On the other hand, humans kill over 100 million sharks annually.

[12] An application of very different nature played an important role in World War II. To plan the location of key infrastructure, British intelligence analysed the distribution of German missile impacts in certain territories and noticed that hits were well approximated by the Poisson distribution. From this, they could deduce that bombs struck at random and could not be aimed accurately. For details, see Example 2.24 in *Elementary probability for applications* by Durrett.

year? As failures are rare events and we expect three each year, we model the number of failures in one year by a Poisson random variable X with $\lambda = 3$. The probability of at least three failures is

$$\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - e^{-3} - 3e^{-3} - \frac{9}{2}e^{-3} = 1 - \frac{17}{2}e^{-3}.$$

Example 3.31 (Vicious horses)**.** The term 'law of small numbers' was coined by Bortkiewicz in 1898, who studied the number of officers killed by horse kicks in 15 Prussian army corps over a twenty year period. Here is the observed historical data:

number of cases	0	1	2	3	4	≥ 5
total number (observed)	144	91	32	11	2	0
frequency (observed)	0.514	0.325	0.114	0.039	0.007	0
total number (theoretical)	141.3	96.6	33.05	7.6	1.3	0.2
frequency (theoretical)	0.505	0.345	0.118	0.027	0.005	0.001

The theoretical values correspond to the expected frequencies in 280 independent realisations of a random variable X with Poisson distribution with $\lambda = 0.684$ obtained as the sample average. We could also estimate λ from $e^{-\lambda} = \mathbb{P}(X = 0) \approx 0.514$ leading to $\lambda = 0.665$. Asymptotic frequencies approximate the values $\mathrm{Poi}_\lambda(k), k \in \{0, 1, \ldots\}$ by the law of large numbers discussed in Section 6.

Example 3.32 (Radioactive decay)**.** In radioactive material, a large number of nuclides decay over time. Over short intervals, nuclei decay with very small probability, (approximately) independently of each other. In a famous experiment, Rutherford, Chadwick and Ellis [1920] counted the number of α-particles emitted by a radioactive Polonium source in 2608 intervals of 7.5 seconds. Here's the data:

number of α-particles	0	1	2	3	4	5	6	7	≥ 8
total number (obs.)	57	203	383	525	532	408	273	139	88
frequency (obs.)	0.022	0.078	0.147	0.201	0.204	0.156	0.105	0.053	0.034
total number (theor.)	54	211	407	525	508	394	254	140	115
frequency (theor.)	0.021	0.081	0.156	0.202	0.195	0.151	0.097	0.054	0.044

As in the previous example, the theoretical values stem from the Poisson distribution with $\lambda = 3.87$.

3.6 Independent random variables

Given random variables $X_1, \ldots, X_n$ with $X_i : \Omega \to \mathbb{R}$ for $i \in \{1, \ldots, n\}$ and sets $A_1, \ldots, A_n \subseteq \mathbb{R}$, let

$$\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} := \{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \ldots \cap \{X_n \in A_n\}.^{13}$$

[13] You should think of the *comma* in the left hand expression replacing *and*.

Definition 3.33 (Independence of random variables)**.** Let $X_1, \dots, X_n$ be discrete random variables with $X_i : \Omega \to \mathbb{R}$ for $i \in \{1, \dots, n\}$. We say that $X_1, \dots, X_n$ are *independent* if for any elements $x_1 \in S_{X_1}, x_2 \in S_{X_2}, \dots, x_n \in S_{X_n}$ we have

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i).$$

Example 3.34 (Dice)**.** Roll fair dice ⚄ ⚃ and set $\Omega = \{1, \dots, 6\}^2$. Then, $X(\omega) = X((\omega_1, \omega_2)) = \omega_1$ and $Y(\omega) = Y((\omega_1, \omega_2)) = \omega_2$ give the outcomes in the individual rolls (X for ⚄ and Y for ⚃). Note that $S_X = S_Y = \{1, \dots, 6\}$. Using the uniform distribution $\mathbb{P}$ on Ω gives, for $x, y \in \{1, \dots, 6\}$,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}\big(\{(\omega_1, \omega_2) \in \Omega : X((\omega_1, \omega_2)) = x, Y((\omega_1, \omega_2)) = y\}\big) = \mathbb{P}\big(\{(x, y)\}\big) = \frac{1}{36}.$$

Similarly, for $x \in \{1, \dots, 6\}$,

$$\mathbb{P}(X = x) = \mathbb{P}\big(\{(\omega_1, \omega_2) \in \Omega : X\big((\omega_1, \omega_2)\big) = x\}\big) = \mathbb{P}\big(\{(x, \omega_2) : \omega_2 \in \{1, \dots, 6\}\}\big) = \frac{6}{36} = \frac{1}{6}.$$

In the same way, one proves $\mathbb{P}(Y = y) = 1/6$ for all $y \in \{1, \dots, 6\}$. Hence, X, Y are independent.

Example 3.35 (Dice)**.** We roll fair dice ⚅ ⚂ repeatedly until they show different numbers. We can use $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, \dots, 6\}, \omega_1 \neq \omega_2\}$ to model the experiment and take uniform $\mathbb{P}$ on Ω. As before, set $X(\omega) = X((\omega_1, \omega_2)) = \omega_1$ and $Y(\omega) = Y((\omega_1, \omega_2)) = \omega_2$. For $x \in \{1, \dots, 6\}$, we have

$$\mathbb{P}(X = x) = \mathbb{P}(\{(\omega_1, \omega_2) \in \Omega : X((\omega_1, \omega_2)) = x) = \mathbb{P}(\{(x, \omega_2) : \omega_2 \in \{1, \dots, 6\} \setminus \{x\}\}) = \frac{5}{30} = \frac{1}{6},$$

and, analogously, $\mathbb{P}(Y = y) = 1/6$ for all $y \in \{1, \dots, 6\}$. On the other hand, we have $\mathbb{P}(X = x, Y = x) = 0$ for all $x \in \{1, \dots, 6\}$. Hence, X and Y are not independent.

Example 3.36. Let X, Y be two independent random variables with the uniform distribution on $S_X = S_Y = \{1, \dots, 6\}$. That is, $\mathbb{P}(X = k) = \mathbb{P}(Y = k) = 1/6$, for $k = 1, \dots, 6$. What is the mass function of the random variable $Z = |X - Y|$? First of all, $S_Z = \{0, \dots, 5\}$. Then,

$$\mathbb{P}(Z = 0) = \mathbb{P}\left(\bigcup_{i=1}^{6} \{X = i, Y = i\}\right) = \sum_{i=1}^{6} \mathbb{P}(X = i, Y = i) = \sum_{i=1}^{6} \mathbb{P}(X = i)\,\mathbb{P}(Y = i) = \frac{1}{6},$$

where we could replace the probability of the union by the sum of the probabilities as the events are disjoint and split up the probability in the last step since X, Y are independent. Next, for $k = 1, \dots, 5$, by similar arguments,

$$\begin{aligned}
\mathbb{P}(Z = k) = \mathbb{P}(\{X - Y = k\} \cup \{Y - X = k\}) &= 2 \cdot \mathbb{P}(X - Y = k) \\
&= 2 \cdot \mathbb{P}\left(\bigcup_{i=k+1}^{6} \{X = i, Y = i - k\}\right) \\
&= 2 \sum_{i=k+1}^{6} \mathbb{P}(X = i, Y = i - k) \\
&= 2 \sum_{i=k+1}^{6} \mathbb{P}(X = i)\,\mathbb{P}(Y = i - k) = \frac{6 - k}{18}.
\end{aligned}$$

Proposition 3.37. Let $X_1, \ldots, X_n$ be independent random variables. Then, given $A_i \subseteq S_{X_i}$ for $i = 1, \ldots, n$, we have

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right).$$

Proof. Let $X_1, \ldots, X_n$ be independent and $A_i \subseteq S_{X_i}$, $i = 1, \ldots, n$. Then,

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{j=1}^{n} \{X_i \in A_i\}\right) &= \sum_{x_1 \in A_1} \cdots \sum_{x_n \in A_n} \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \sum_{x_1 \in A_1} \cdots \sum_{x_n \in A_n} \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \\
&= \prod_{i=1}^{n} \Big(\sum_{x_i \in A_i} \mathbb{P}\left(X_i = x_i\right) \Big) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right)
\end{aligned}$$

where we used independence in the second step. □

Proposition 3.38. Let $X_1, \ldots, X_n$ be independent random variables and $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be arbitrary functions. Then the random variables $f_1 \circ X_1, \ldots, f_n \circ X_n$ are independent.

For example, this means that if X, Y are independent then so are the pairs X^2, Y^3 or $\sin X, e^Y$.

Proof. Let $y_1, \ldots, y_n \in \mathbb{R}$. Set $A_i = \{x \in \mathbb{R} : f(x) = y_i\}$. Then,

$$\mathbb{P}\big(f_1 \circ X_1 = y_1, \ldots, f_n \circ X_n = y_n\big) = \mathbb{P}\big(X_1 \in A_1, \ldots, X_n \in A_n\big) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right) = \prod_{i=1}^{n} \mathbb{P}\left(f_i \circ X_i = y_i\right)$$

where we used independence of $X_1, \ldots, X_n$ and Proposition 3.37 in the second step. This shows that $f_1 \circ X_1, \ldots, f_n \circ X_n$ are independent. □

Proposition 3.39. Let X, Y be independent discrete random variables and $Z = X + Y$. Then, $S_Z = \{x + y : x \in S_X, y \in S_Y\}$. For all $z \in S_Z$, we have

$$\mathbb{P}\left(X + Y = z\right) = \sum_{x \in S_X} \mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = z - x\right).$$

Proof. Let $z \in S_Z$. Then we can write $\mathbb{P}\big(X + Y = z\big)$ as

$$\mathbb{P}\Big(\bigcup_{x \in S_X} \{X = x, Y = z - x\} \Big) = \sum_{x \in S_X} \mathbb{P}\left(X = x, Y = z - x\right) = \sum_{x \in S_X} \mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = z - x\right).$$

The second equality here uses that the events $\{X = x, Y = z - x\}, x \in S_X$ are pairwise disjoint and the third equality follows from the independence of X, Y. □

We note that independence is crucial in Proposition 3.39. Using this result one can show the following.

Proposition 3.40. Let X, Y be independent random variables.

(i) If $X \sim \text{bin}_{n,p}$ and $Y \sim \text{bin}_{m,p}$ for $n, m \geq 1$ and $0 \leq p \leq 1$, then $X + Y \sim \text{bin}_{n+m,p}$.

(ii) If $X \sim \text{Poi}_\lambda$ and $Y \sim \text{Poi}_\mu$ for $\lambda, \mu > 0$, then $X + Y \sim \text{Poi}_{\lambda+\mu}$.

Most important takeaways in this section.

You should

- understand the definition of a discrete random variable,
- be able to compute mass and distribution functions in simple cases and know their properties,
- be familiar with factorials and binomial coefficients and their applications to counting,
- know mass functions of binomial, hypergeometric, geometric and Poisson distributions and be able to use these in standard scenarios,
- be able to apply the law of small numbers and to formulate it correctly,
- understand the concept of independence of random variables and be able to check it.

Appendix

Sketch proof of Theorem 3.26. To begin fix $k \in \{0, 1, 2, ...\}$ as in the statement. For each $n \in \mathbb{N}$, we set $p_n = \lambda_n/n$. From the statement of the theorem, we have $\lambda_n \to \lambda$ as $n \to \infty$. Therefore

$$\begin{aligned}\text{bin}_{n,p_n}(k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!}\Big(\frac{\lambda_n}{n}\Big)^k\Big(1-\frac{\lambda_n}{n}\Big)^{n-k} \\ &= \Big(\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\Big)\cdot\Big(1-\frac{\lambda_n}{n}\Big)^{-k}\cdot\frac{(\lambda_n)^k}{k!}\Big(1-\frac{\lambda_n}{n}\Big)^n \\ &\longrightarrow 1\cdot 1\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda},\end{aligned}$$

as $n \to \infty$. To justify the final step, <u>as k is fixed</u>, note that:

- For each $i \in \{0, \ldots, k-1\}$ we have $(n-i)/n \to 1$ as $n \to \infty$.
- As $\lambda_n \to \lambda$ we have $(1 - \lambda_n/n)^{-k} \to 1$ as $n \to \infty$.
- As $\lambda_n \to \lambda$ we have $(\lambda_n)^k \to \lambda^k$.
- In Real Analysis it was proved that for all $\lambda \in \mathbb{R}$ we have $(1 - \lambda/n)^n \to e^{-\lambda}$. We obtain that $\lim_{n\to\infty}(1 - \lambda_n/n)^n = e^{-\lambda}$ for a similar reason [14].

By the Algebra of Limits (AoL), these combine to give the limit claimed above. □

[14] If you would like, you can use the inequalities $e^{x-x^2} \leq 1 + x \leq e^x$ for $x \in [-1/2, 0]$ from Section 1.4.3 to prove this, taking $x = -\lambda_n/n$.