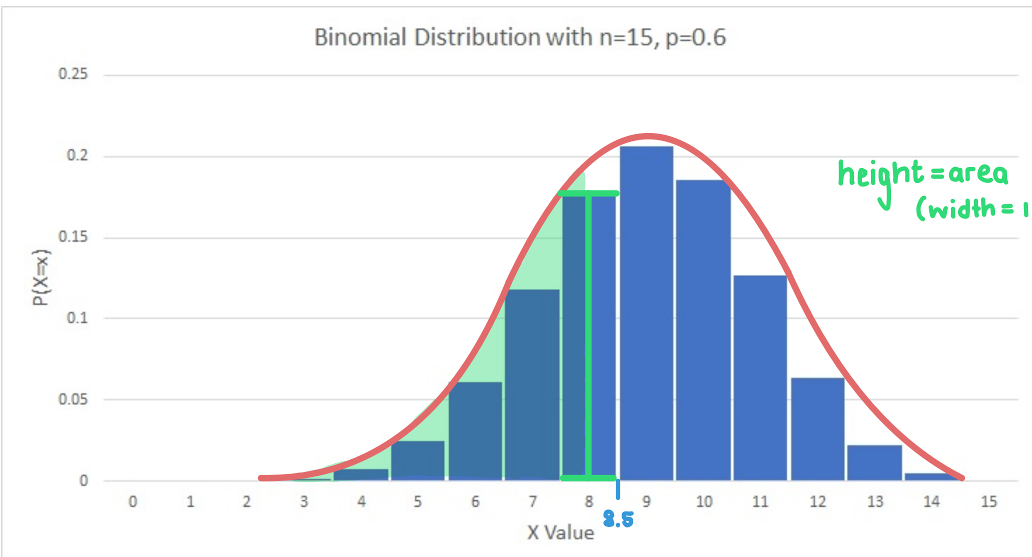
Stat 260 Lecture Notes
Set 16 - The Normal Approximation of the Binomial Distribution

Consider a binomial experiment with n trials where the probability of success is p.
Sometimes the pmf (pictured as a bar chart) of the binomial distribution has a bell-curve shape.

Example 1: Consider the binomial distribution with $n = 15$ and $p = 0.60$. This distribution is pictured below.



The pmf pictured here looks approximated bell-curved shape, so we can use the normal distribution to approximate our probability calculations.

Let's work through the calculation of $P(X \leq 8)$.

$P(X=0) + P(X=1) + P(X=2) + \cdots + P(X=7) + P(X=8)$

Recall: In a binomial distribution we have $E(X) = \mu = np$, and $V(X) = \sigma^2 = np(1-p)$, and $\sigma = \sqrt{np(1-p)}$.

This tells us that if we want to standardize the random variable X in the normal distribution we would use

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

There is one complicating factor though. The bars of the bar chart have width. So comparing the area of $P(X \leq 8)$ in the binomial bars will be slightly different from the area of $P(X \leq 8)$ in the normal distribution curve. Instead, we have to do a bit of a correction to our normal distribution. Notice that the X=8 bar actually ends at the value X=8.5. So when calculating the area in the normal distribution we should look at $P(X \leq 8.5)$.

If we want $P(X \leq x)$ in the binomial distribution we will use $P(X \leq x) \approx P(X \leq x + 0.5)$ in the normal distribution. This means that our standardization will instead become

$$Z = \frac{X - \mu}{\sigma} = \frac{X + 0.5 - np}{\sqrt{np(1-p)}}$$

The value of the $+0.5$ is called the **continuity correction factor**.

Notice that we can calculate $P(X \geq x)$, $P(X > x)$, $P(X < x)$, and $P(X = x)$ too by first changing the statement to involve $P(X \leq x)$.

When can we use the normal approximation to the binomial distribution?

Rule: The normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ is a good approximation to the binomial(n, p) when $np \geq 5$ and $n(1-p) \geq 5$.

Example 1: Look at the binomial distribution with $n = 15$ and $p = 0.60$. Use the normal approximation to the binomial distribution to calculate:

(a) $P(X \leq 8)$

(b) $P(X = 8)$

(c) $P(X \geq 5)$

Compare each approximation to the exact value you would find with the binomial distribution.

$$\mu = np = 15(0.6) = 9 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{15(0.6)(0.4)} = \sqrt{3.6}$$

a) $P(X \leq 8) \approx P(X \leq 8.5)$

$$z = \frac{x-\mu}{\sigma} = \frac{8.5-9}{\sqrt{3.6}} = -0.26$$

$\therefore P(z \leq -0.26) = 0.3974$

get from table

Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

b) $P(X=8) = P(X \leq 8) - P(X \leq 7) \approx P(X \leq 8.5) - P(X \leq 7.5)$

$$z = \frac{x-\mu}{\sigma} = \frac{8.5-9}{\sqrt{3.6}} = -0.26 \qquad z = \frac{x-\mu}{\sigma} = \frac{7.5-9}{\sqrt{3.6}} = -0.79$$

Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

$\therefore P(z \leq -0.26) - P(z \leq -0.79)$
$= 0.3974 - 0.2148$
$= 0.1826$

c) $P(X \geq 5) = 1 - P(X \leq 4) \approx 1 - P(X \leq 4.5)$

$$z = \frac{x-\mu}{\sigma} = \frac{4.5-9}{\sqrt{3.6}} = -2.37$$

Table A.3 Areas under the Normal Curve

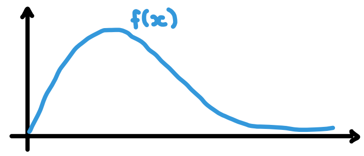
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183

$\therefore 1 - P(z \leq -2.37)$
$= 1 - 0.0089$
$= 0.9911$

Stat 260 Lecture Notes
Set 17 - The Gamma Distribution and Exponential Distribution

The **gamma distribution** is used to model right-skewed continuous data.

The r.v. X is gamma distributed, if it has pdf



$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

for $x > 0$ where α and β are parameters, and $\Gamma(\alpha)$ is the **gamma function**. (Note: Sometimes the distribution is described in terms of k and θ instead of α and β. In this setup $k = \alpha$ and $\theta = \frac{1}{\beta}$.)

If X is gamma distributed we write $X \sim \text{gamma}(\alpha, \beta)$.

The **gamma function** $\Gamma(\alpha)$ is defined as

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}\, dx$$

Rules:

- If $X \sim \text{gamma}(\alpha, \beta)$, then $E(X) = \alpha\beta$ and $V(X) = \alpha\beta^2$.
- $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$
- $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- $\Gamma(n) = (n-1)!$ for positive integers n

Example 1: Say that the time it takes to write a stat midterm is gamma distributed with $\alpha = 2$ and $\beta = 20$. What is the probability that a random student writing the midterm will finish in under 47 minutes?

The r.v. X follows the **exponential distribution** with parameter λ ($\lambda > 0$) if the pdf is

$$f(x) = \lambda e^{-\lambda x}$$

and here $x \geq 0$.

We can find $E(X)$ by calculating $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}$.

We can find $V(X)$ by calculating $V(X) = E(X^2) - [E(X)]^2$
$= \int_0^{\infty} x^2\lambda e^{-\lambda x}\, dx - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$.
(This is found by doing integration by parts twice.)

So for the exponential distribution we have

$$E(X) = \mu = \frac{1}{\lambda} \text{ and } V(X) = \sigma^2 = \frac{1}{\lambda^2}$$

We can find the cdf $F(x)$ for the exponential distribution by:

$$F(y) = \int_0^x f(y)\, dy$$
$$= \int_0^x \lambda e^{-\lambda y}\, dy$$
$$= -e^{-\lambda y}\Big|_0^x$$
$$= -e^{-\lambda x} + 1$$

Note: The exponential distribution is a special case of the gamma distribution where $\alpha = 1$ and $\beta = \frac{1}{\lambda}$.

Example 2: Suppose the length of a customer service call in a call center (measured in minutes) is an exponential random variable with parameter $\lambda = \frac{1}{10}$. Suppose a worker just answered a call. What is the probability this call will last more than 10 minutes?

$$P(X > 10) = 1 - P(X \leq 10)$$
$$= 1 - 0.3679$$
$$= 0.6321$$

$$f(10) = \frac{1}{10} e^{-\frac{1}{10}(10)}$$
$$= 0.3679$$

What is the probability this call will last between 10 and 20 minutes?

$$P(10 \leq X \leq 20) = P(X \leq 20) - P(X \leq 10)$$
$$= e^{-1} - e^{-2}$$
$$= 0.2325$$

A summary of three types of similar sounding random variables:

- A binomial random variable X counts the number of successes in a **fixed number of trials** n.
- A Poisson random variable X counts the **number of successes in an interval** of time/length/space/etc.
- An exponential random variable X counts the **amount of time between successes** in the Poisson process.

The exponential distribution is related to the Poisson distribution.

Rule: If X is a Poisson random variable with parameter λt (where λ is the average number of events in one unit of time, and t is the number of units of time in the interval of interest), then the distribution of time between occurrences of two events is exponential with parameter λ.
In other words: The λ that we use in the Poisson distribution setup for one unit of time is equal to the λ we use in the exponential distribution setup.

Example 3: Suppose the number of students that email Michelle each day is a Poisson random variable where on average she receives 3 emails per day.
If Michelle just received an email, what is the probability that she will wait more than 1 day until the next email?

Recall:

- For a Poisson random variable X, $E(X) = \mu = \lambda$ and $V(X) = \sigma^2 = \lambda$.
- For an exponential random variable X, $E(X) = \mu = \frac{1}{\lambda}$ and $V(X) = \sigma^2 = \frac{1}{\lambda^2}$.

The exponential distribution has the **memoryless property**, that $P(X \geq a+b \mid X \geq a) = P(X \geq b)$. That is, if we know that an amount of time a has already passed and we want to see the probability that the next success takes a total amount of time at least $a+b$, this is the same as saying after time a has passed, call that marker as time 0 then count the probability of at least a time of b from there.

We can see this by the calculation:

Note: The memoryless property is not the same thing as saying the events "$X \geq a+b$" and "$X \geq b$" are independent. If the events were independent we would have $P(X \geq a+b \mid X \geq a) = P(X \geq a+b)$.