STAT 345 ⋄ THE NORMAL DISTRIBUTION

KELLIN RUMSEY ⋄ SPRING 2020

The normal distribution is one of the most important distributions in all of statistics. In fact, one could argue that the normal distribution is sometimes used more often than it should be. The normal distribution is also know as a *Gaussian* distribtuion or the *Bell-curve* distribution. The normal distribution is characterized by two parameters, the mean μ and the standard deviation σ (or equivalently, the variance σ^2). The distribution is always symmetric, unimodal and "bell-shaped".

Definition: If X has a *normal distribution* with mean μ and sd σ, then the probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \ \infty < x < \infty$$

The expected value and variance of X is

$$E(X) = \mu \qquad \text{Var}(X) = \sigma^2$$

Notationally, we write $X \sim N(\mu, \sigma^2)$.

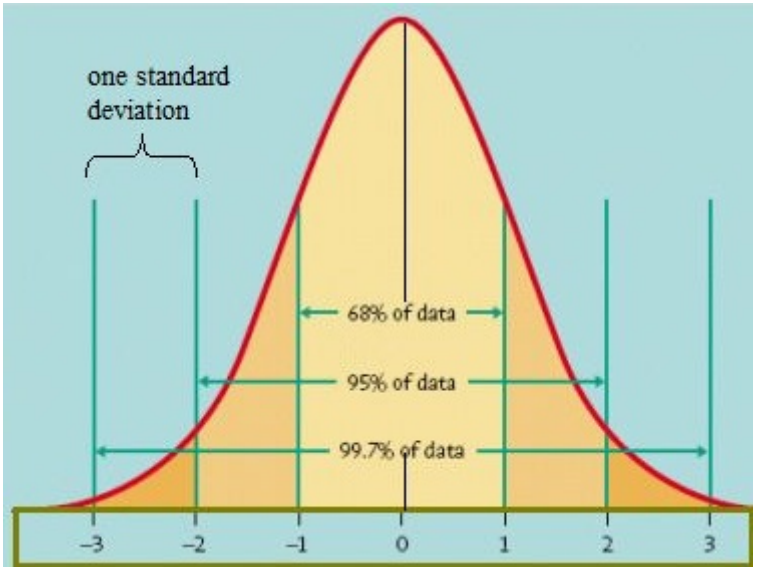

FIGURE 1. Illustration of the so-called $68 - 95 - 99.7$ rule.

The CDF for a normal distribution cannot be written down analytically, and must be computed numerically. We will discuss how to find normal probabilities shortly, but also note that the CDF of a normally distributed RV X can be computed in R as $P(X \leq x) \stackrel{R}{=}$ `pnorm(x, mu, sigma)`.

The $68 - 95 - 99.7$ rule is a very crude way to approximate probabilities for a normal RV. This rule, illustrated in Figure 1, states that

$$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$$

Example: Suppose that the weights of adult male African elephants are normally distributed with a mean of 13 thousand pounds and a standard deviation of 2.5 thousand pounds.

i) What is the probability that a randomly selected African elephant (adult male) weighs between 8 and 18 thousand pounds?

According to the normal distribution, this probability is approximately 0.95 *because* $\mu - 2\sigma = 13 - 2(2.5) = 8$ *and* $\mu + 2\sigma = 13 + 2(2.5) = 18$.

ii) What is the probability that a randomly selected adult male African elephant weighs less than 8 thousand pounds OR more than 18 thousand pounds?

Clearly this is the complement of the event from the previous problem, so the probability is approximately 0.05.

iii) What is the probability that an adult male African elephant weights more than 18 thousand pounds?

Since the normal distribution is symmetric, the probability found in the previous example must be split evenly between the "less than 8 *thousand pounds" and "more than* 18 *thousand pounds" cases. So the probability is approximately* 0.025.

iv) Suppose that Jumbo is an adult male African elephant. About 84% of other adult male African elephants weigh more than Jumbo. How much (approximately) does Jumbo weight?

Let x *be Jumbo's weight and let* X *be the weight of a randomly selected adult male African elephant. The problem statement tells us that* $P(X > x) = 0.84$ *or equivalently that* $P(X \leq x) = 0.16$. *Using similar logic as in the previous problems, we can reason that Jumbo's height must be* $\mu - \sigma = 13 - 2.5 = 11.5$ *thousand pounds.*

Next, we state an important fact about the normal distribution.

Theorem 1: If X is normally distributed, then so is any linear function of X. In other words, for constants a and b, the RV $aX + b$ is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$.

Definition: We say that Z has a *standard normal distribution* if it is normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 1$. I.e. $Z \sim N(0, 1)$.

The PDF of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The CDF of Z is very important (at least historically) in statistics, so rather than using the usual $F(z)$, we use a capital "phi" to denote it. That is,

$$P(Z \leq z) = \Phi(z).$$

The CDF $\Phi(z)$ still must be computed numerically, but the value can be found for a large number of z using a Standard Normal Table. This was very important/useful in the days before computing power was readily available. In this class, we will try to use R (or the internet) to calculated the normal CDF, but it is fairly straightforward to use the tables when required (demonstration was done in class).

Example. Let Z be a standard normal random variable. Find the following probabilities.

i)
$$P(Z \leq 1.45) = \Phi(1.45) = 0.9265$$

ii)
$$P(Z > 0.55) = 1 - P(Z \leq 0.55) = 1 - \Phi(0.55) = 1 - 0.7088 = 0.2912$$

iii)
$$P(-1.36 < Z < 1.45) = \Phi(1.45) - \Phi(-1.36) = 0.9265 - 0.0869 = 0.8396$$

Alternatively, I could give you a probability statement and ask you to find the corresponding z. For example, if $P(Z > z) = 0.25$, then what is the value of z?

$$P(Z > z) = 0.25$$
$$1 - P(Z \leq z) = 0.25$$
$$\Phi(z) = 0.75$$

This means that $z = \Phi^{-1}(0.75)$. This value can be found by "reversing the process" when using the table (demonstrated in class), or in R using the "quantile function". In R,

$$\Phi^{-1}(p) \stackrel{R}{=} \texttt{qnorm(p, mu, sigma)}$$

For this example, we find that z must be equal to $z = 0.67$.

Standardization. The nice thing about Theorem 1, is that we can use it to go back and forth between any two normal distributions. The most common cases are given now.

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ has a standard normal distribution.

If $Z \sim N(0, 1)$, then $X = \mu + \sigma z$ has a $N(\mu, \sigma^2)$ distribution.

This means that we can express the CDF of a normal distribution (for any μ and σ) in terms of the standard normal CDF.

$$\begin{aligned} P(X \leq x) &= P(X - \mu \leq x - \mu) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Example: Let X be normally distributed with mean $\mu = 70$ and sd $\sigma = 3$.

i) What is the probability that X is at least 74?

$$P(X > 74) = 1 - P(X \leq 74) = 1 - \Phi\left(\frac{74-70}{3}\right) = 1 - \Phi(1.33) = 1 - 0.9082 = 0.0917$$

ii) What is the probability that X is between 70 and 74?

$$P(70 < X < 74) = P(X \leq 74) - P(X \leq 70) = \Phi\left(\frac{74-70}{3}\right) - \Phi\left(\frac{70-70}{3}\right) = 0.9082 - \Phi(0) = 0.9082 - 0.5000 = 0.4082$$

i) What value x is exceeded with probability 0.6?

$$P(X > x) = 0.6 \quad \Rightarrow \quad P(X \leq x) = 0.4 \quad \Rightarrow$$

$$\Phi\left(\frac{x-70}{3}\right) = 0.4 \quad \Rightarrow \quad \frac{x-70}{3} = \Phi^{-1}(0.4) \quad \Rightarrow$$

$$x = 70 + 3\Phi^{-1}(0.4) = 70 + 3(-0.25) = 69.25$$

Normal Approximations. By typing "Binomial distribution applet" into google, you can find a tool which will plot the PMF of a Binomial distribution for any values of n and p. Note that when n is large (and p is not too close to 0 or 1), the distribution looks very much like a normal distribution.

Remember that np is the "expected" number of successes in the experiment. You are probably not surprised to hear that $n(1-p)$ is the "expected" number of failures in the experiment. As long as both of these quantities are reasonably large (some textbooks suggest they should be at least 15) then the normal approximation is quite good.

This gives us an easy way of approximating probabilities which are otherwise too tedious (by hand) or too computationally difficult (n has to be very very large for this to be a problem). The rough idea, is that the Binomial CDF can be approximated using the normal CDF. Let $X \sim Binom(n, p)$, then

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) \qquad \text{(not the correct formula).}$$

Similarly, we could define

$$P(X \geq x) \approx 1 - \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) \qquad \text{(not the correct formula).}$$

There is a small problem with these results. Since the Binomial distribution is discrete, these formulas won't always match. For example, let $n = 100$ and $p = 0.5$, so that $\mu = np = 50$ and $\sigma = \sqrt{np(1-p)} = 5$, and suppose we are looking for $P(X < 40)$. There are two, seemingly correct ways to approximate this probability.

$$P(X < 40) = P(X \leq 39) \approx \Phi\left(\frac{39 - 50}{5}\right) = \Phi(-2.2) = 0.0139$$

$$P(X < 40) = 1 - P(X \geq 40) \approx 1 - \left(1 - \Phi\left(\frac{40 - 50}{5}\right)\right) = \Phi(-2.0) = 0.0227$$

Do you see the problem? We can fix this by applying a *continuity correction*, which typically leads to a better approximation as well. Here is the final result written down.

Normal approximation to the Binomial. Let X have a binomial distribution with n trials and success probability p. If np and $n(1-p)$ are large, then X is approximately normally distributed with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$, and the following approximations hold.

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) \qquad P(X \geq x) \approx 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

You may remember that there is a relationship between the Binomial distribution and the Poisson distribution. So it should be unsurprising that a similar result holds for the Poisson distribution when the expected value λ is large. For the same reasons as before, we will apply a continuity correction.

Normal approximation to the Poisson. Let X have a Poisson distribution with parameter λ. If λ is large, then X is approximately normally distributed with mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$, and the following approximations hold.

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - \lambda}{\sqrt{\lambda}}\right) \qquad P(X \geq x) \approx 1 - \Phi\left(\frac{x - 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

These formulas should be fairly straightforward to apply. For examples, see the homework questions about these topics (and ask me if you have any questions).

The Lognormal Distribution. There are many real life applications, where the variable of interest must take on a positive value. Examples include time, weight, height, mass, etc. The normal distribution may not be a good choice for many of these variables, especially those with heavy skew. The exponential and uniform distributions are often not flexible enough to be a good choice either. The Weibull, Gamma and Lognormal distributions are more flexible distributions which can be used to describe the distribution for many real world applications. Weibull was covered very briefly in the previous lecture notes. Gamma will be covered briefly after spring break. Here, we briefly cover the Lognormal distribution.

Definition. Let Y be normally distributed with mean θ and standard deviation ω. If $X = e^Y$, then we say that X has a *lognormal* distribution with "log-mean" θ and log-sd ω. The PDF of X can be written down (look it up online if you want) but is not very useful. Notationally, we write $X \sim logN(\theta, \omega)$ and we have

$$E(X) = e^{\theta + \omega^2/2} \qquad Var(X) = \left(e^{\omega^2} - 1\right) e^{2\theta + \omega^2}$$

The CDF can be written as

$$P(X \leq x) = \Phi\left(\frac{\ln(x) - \theta}{\omega}\right), x > 0$$

Other facts about the lognormal distribution include that its median is e^θ and its skewness is

$$\text{Skew}(X) = \left(e^{\omega^2} + 2\right)\sqrt{e^{\omega^2} - 1}.$$

Note that the skewness is always positive. It is fairly close to 0 (i.e. symmetric) when ω is very small. When ω is large however the skew can be very large, indicating that this distribution is useful for modeling data with a heavy right skew. The last section of these notes focuses on further explaining why the lognormal distribution gets its name, and deriving the CDF given above.

If $X \sim logN(\theta, \omega)$ and $Y = \ln(X)$, then Y is normally distributed with mean θ and variance ω^2. Hence the name, taking the log leads to a normal distribution. The CDF of the lognormal distribution can be derived as:

$$\begin{aligned} P(X \leq x) &= P(\ln(X) \leq \ln(x)) \\ &= P(Y \leq \ln(x)) \\ &= P\left(\frac{Y-\theta}{\omega} \leq \frac{\ln(x)-\theta}{\omega}\right) \\ &= \Phi\left(\frac{\ln(x)-\theta}{\omega}\right) \end{aligned}$$

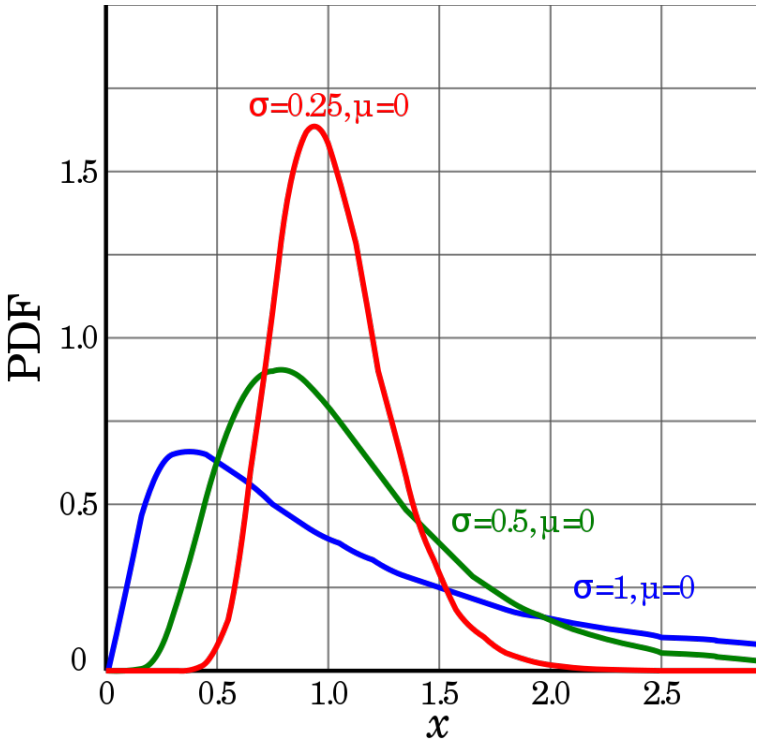

FIGURE 2. The lognormal distribution for 3 different sets of parameters.