The Normal Distribution and the Central Limit Theorem

The Normal Distribution

The **normal distribution** (also called the **Gaussian distribution**) is by far the most commonly used distribution in statistics. It models a continuous random variable whose density takes a bell-curve shape. You can see this curve in Figure 1 below.

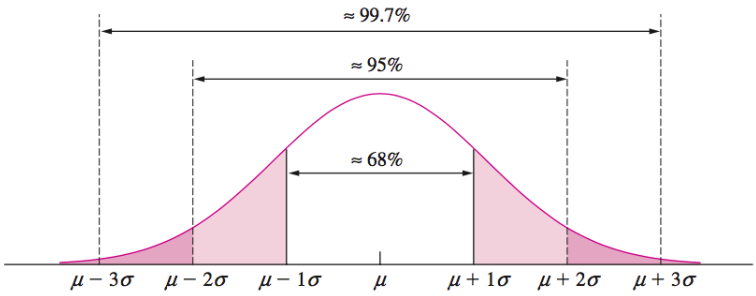

Figure 1: Density shape of a normal random variable with mean μ and variance σ^2.

When a random variable X is normally distributed, we write $X \sim N(\mu, \sigma^2)$, and its probability density function is[1]

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

Properties and Standard Units

The normal distribution PDF may seem complicated, but the curve has several nice properties that are depicted in Figure 1:

[1] There is a typo in the textbook: they forgot the negative sign in the exponential.

For one, it is completely specified in terms of its mean μ and variance σ^2. These are the only two properties that distinguish different Gaussian curves.

Moreover, the curve is *symmetric* around μ, so μ is both the mean and the median.

Finally, the proportion of density within a given number of standard deviations of μ will be the same for any normal population. Roughly 68% will be within $\mu \pm \sigma$, 95% within $\mu \pm 2\sigma$, and 99.7% within $\mu \pm 3\sigma$. This will always hold, regardless of what μ and σ are.

Because of this last point, when dealing with normal populations, we often convert the data to **standard units**, which tell how many standard deviations an observation is from the population mean. The standard units follow $Z \sim N(0, 1)$, and we convert a datapoint in X to one in Z by

$$z = \frac{x - \mu}{\sigma}$$

This value is sometimes called the "z-score" of x.

Area Under the Curve

How do we find the proportion of a normal population within a given interval? To find the area analytically, we would need to integrate $f(x)$ between the two points. But trying to do this with the Normal distribution is actually impossible, because its antiderivative is an infinite series.

Instead, we must rely on numerical approximations. Areas under the standard normal curve are very commonly tabulated in tables called the **standard normal table**, or **z table** (table A.2 in the textbook). These tables give the cumulative probability at different points, *i.e.*, $P(Z \leq z)$ for different values of z, and we can use these to approximate the areas.

Let's consider an example from the book. A process manufactures ball bearings whose diameters are normally distributed with mean 2.505 cm and standard deviation 0.008 cm. Specifications call for the diameter to be in the interval 2.5 ± 0.01 cm. What proportion of the ball bearings will meet the specification?

If X is the random variable for diameter, then $X \sim N(2.505, 0.008^2)$. We can draw the PDF as in Figure 2:[2] the curve is symmetric about $\mu = 2.505$,

[2] By hand, we wouldn't worry about the exact y-values.

and we want to know the probability that X is between 2.49 and 2.51 cm, *i.e.*, $P(2.49 < X < 2.51)$. To do this, we find the z-scores of 2.49 and 2.51:

$$z = \frac{2.49 - 2.505}{0.008} = -1.88 \qquad z = \frac{2.51 - 2.505}{0.008} = 0.63$$

Then we look up the areas in the z-table: $P(Z < -1.88) = 0.0301$ and $P(Z < 0.63) = 0.7357$. Putting these together, $P(-1.88 < Z < 0.63) = 0.7357 - 0.0301 = 0.7056$. So, 70.6% of the ball bearings meet the spec.

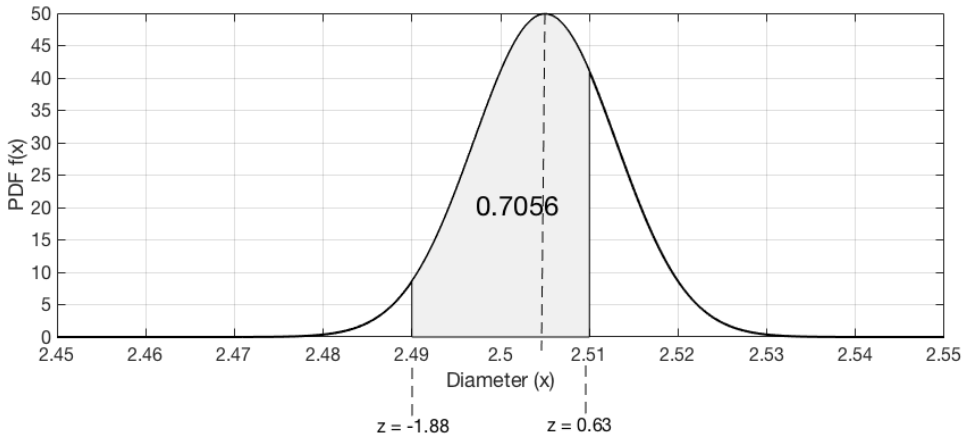

Figure 2: Graph of the bell curve for this example.

If we re-calibrated the process so that the mean is at the center of the specification interval (*i.e.*, $\mu = 2.5$), do we expect the proportion to increase or decrease? It would increase, because the normal distribution has the highest density centered around its mean.

What if we wanted to increase the proportion in spec by some specified amount? Adjusting the mean can only bring us so far – we would have to fine-tune the standard deviation to get additional improvements.

How could we increase it to 90%? We need $P(2.49 < X < 2.51) = 0.9$. There are many ways to do this, but if we want to adjust σ by the smallest amount, let's assume we first adjust $\mu = 2.50$, as shown in Figure 3. Then, we can look for the z-scores with 5% of the area below and 5% above: these are roughly $z = -1.645$ and $z = +1.645$, *i.e.*, $P(-1.645 < Z < +1.645) = 0.90$. We can use either of these values to find σ by determining what it must be so the bounds of X correspond to these z-scores:

$$-1.645 = \frac{2.49 - 2.5}{\sigma} \qquad \text{or} \qquad 1.645 = \frac{2.51 - 2.5}{\sigma}$$

Solving gives $\sigma = 0.0061$, which is about a 24% reduction from what it was before.

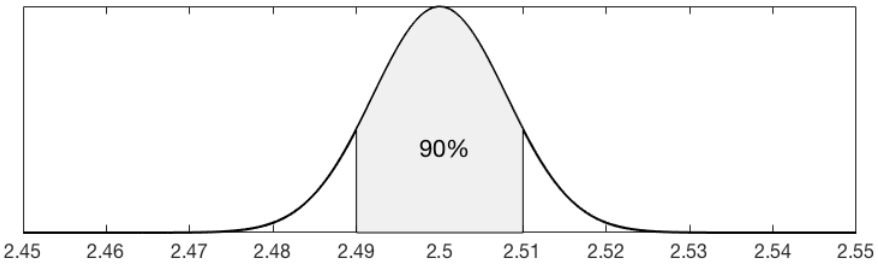

Figure 3: We need σ such that 90% of the area is between 2.49 and 2.51.

Parameter Estimation

If we want to fit a population to a normal distribution, we first draw a random sample $X_1, ..., X_n$. We then estimate μ with the sample mean $\bar{X}$ and σ^2 with the sample variance s^2, giving us $N(\mu, \sigma^2)$. The estimate of μ is unbiased, and has uncertainty $\sigma/\sqrt{n}$.

Linear Functions and Linear Combinations

Sometimes, we need to take a linear function of a random variable X. If X is normal, then any linear function will also be normal. Specifically, if $X \sim N(\mu, \sigma^2)$ and $a \neq 0$, b are constants, then

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Other times, we want to combine several random variables together. One of the remarkable features of the Gaussian distribution is that linear combinations of independent normal random variables are themselves normal random variables. In particular, if $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2), ..., X_n \sim N(\mu_n, \sigma_n^2)$ are n independent, normal random variables, then for constants $c_1, c_2, ..., c_n$

$$\sum_i c_i X_i \sim N\left(\sum_i c_i \mu_i,\ \sum_i c_i^2 \sigma_i^2\right)$$

There are two important, special cases of this result. The first is if $X_1, ..., X_n$ are samples from the same population with mean μ and variance σ^2, and

we average them to generate $\bar{X}$:

$$\bar{X} = (1/n)\sum_i X_i \sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$$

So, the mean stays the same, and the variance shrinks by n. This should remind you of the uncertainty in estimating μ from $\bar{X}$.

The second is for the sum and difference of two random variables: if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then

$$X + Y \sim N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$
$$X - Y \sim N\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$

So, the mean shifts based on whether we add or subtract, while the variance adds in either case.

As an example, suppose $A \sim N(10, 2)$, $B \sim N(-5, 1)$, and $C \sim N(20, 5)$, and we want to know the distribution of $W = 2A+3B+C$. The mean $\mu_W = 2\cdot 10 - 5\cdot 3 + 20\cdot 1 = 25$, and the variance $\sigma_W^2 = 2^2\cdot 2 + 3^2\cdot 1 + 1^2\cdot 5 = 22$, so $W \sim N(25, 22)$. What is *e.g.*, the probability that $30 < W < 40$? We compute the z-scores:

$$z = \frac{30-25}{\sqrt{22}} = 1.07 \qquad\qquad z = \frac{40-25}{\sqrt{22}} = 3.20$$

So, from the z-table,

$$P(30 < W < 40) = P(1.07 < Z < 3.20) = 0.9993 - 0.8577 = 0.1416$$

The Central Limit Theorem

The **Central Limit Theorem** states that if we draw a large enough sample from a population, then the distribution of the sample mean is approximately normal. This holds *no matter what population we are drawing from*: discrete or continuous, skewed or symmetric, etc. When we are averaging large enough sets of numbers, the distribution of the population is irrelevant. It is an extremely surprising, useful and important result in statistics.

In particular, let $X_1, ..., X_n$ be a simple random sample from a population with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$, and let $\bar{X} = S_n/n$ be the sample mean. Then if n is sufficiently large,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{approximately}$$

$$S_n \sim N\left(n\mu, n\sigma^2\right) \qquad \text{approximately}$$

How "large" is large enough? *This* answer depends on the population: the more skewed the distribution, the larger n probably needs to be. However, empirical evidence suggests that for most populations, if $n \geq 30$, the Central Limit Theorem approximation is good.

For example, suppose we know that the mean income of engineers in the US (in thousands) is \$80, with a standard deviation of \$20. Suppose we take a sample of 100 engineers and find the mean of this sample. Since $n = 100 > 30$, the Central Limit Theorem applies, so the distribution of $\bar{X}$ is normal with mean $\mu = 80$ and variance $\sigma^2 = 20^2/100 = 4$, *i.e.*, $\bar{X} \sim N(80, 4)$. Then we can find *e.g.*, the probability that $\bar{X} > 85$ using the z-table:

$$P(\bar{X} > 85) = P(Z > 2.5) = 1 - 0.9938 = 0.0062$$

which is less than 1%.

Normal Approximation to the Binomial

Recall that the binomial distribution $Y \sim \text{Binom}(n, p)$ is the sum of n independent Bernoulli random variables $X_1, ..., X_n$, where each X_i has mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. If n is large enough, then, the Central Limit Theorem tells us that Y can be approximated by a normal distribution. In particular, with $Y = X_1 + \cdots + X_n$ and $\hat{p} = Y/n$ as the estimate of p,

$$Y \sim N\left(np, np(1-p)\right) \qquad \text{approximately}$$

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \text{approximately}$$

So, how "large" is large enough? For the Binomial distribution, it has to do with the mean number of successes np and the mean number of failures $n(1-p)$. The standard rule is that the normal approximation works whenever both np and $n(1-p) > 10$.

Consider this example from the book: In a certain large university, 25% of the students are over 21 years of age. Suppose we take a sample of 400 students, and want to know the probability that a certain number are over 21. If X is the random variable of the number over 21, then $X \sim \text{Binom}(400, 0.25)$. Since $400 \cdot 0.25 = 100 > 10$ and $400 \cdot 0.75 = 300 > 10$, the

Central Limit Theorem holds: $\mu = 400 \cdot 0.25 = 100$ and $\sigma^2 = 400 \cdot 0.25 \cdot 0.75 = 75$, so $X \sim N(100, 75)$. You can see these two distributions plotted together in Figure 4 below; they are very close.

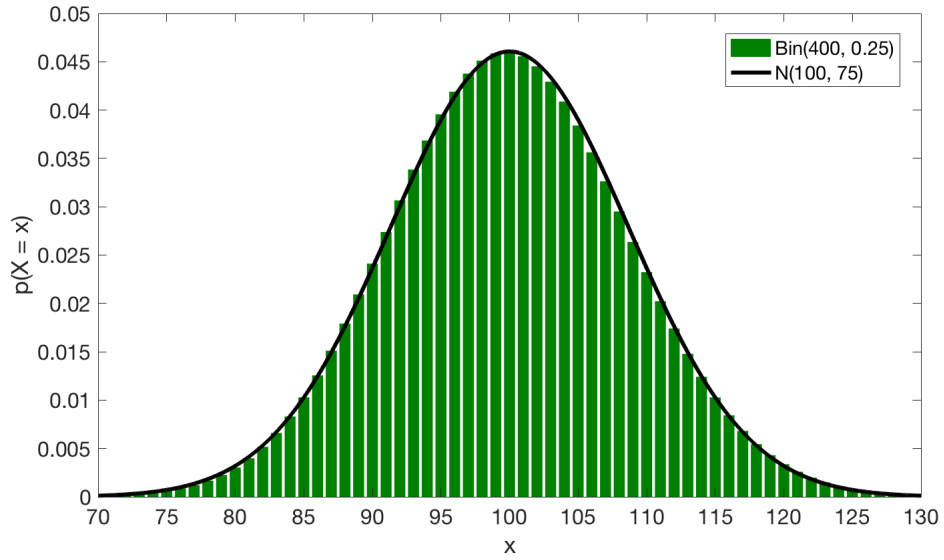

Figure 4: Bin$(400, 0.25)$ and its normal approximation $N(100, 75)$.

What is *e.g.*, the probability that more than 110 of them are over 21? We need $P(X > 110)$. This would be very difficult to do from the Binomial distribution, since we would need to add $P(X = 111) + P(X = 112) + P(X = 113) + \cdots$. Instead, we can find $P(X > 110.5)$ from the normal distribution: $z = (110.5 - 100)/\sqrt{75} = 1.21$, so

$$P(X > 110.5) = P(Z > 1.21) = 1 - 0.8869 = 0.1131$$

Why 110.5 rather than 110? Because we are *excluding* 110: We technically need to start at the rectangle centered at 111, which starts at 110.5 in the continuous approximation.

Normal Approximation to the Poisson

Recall that if $X \sim$ Poisson(λ), then X is approximately binomial with n large and $\lambda = np$, *i.e.*, $X \sim$ Binom(n, λ). When $np > 10$, then, the Central Limit Theorem tells us that X is approximately normal, *i.e.*, $X \sim N(\lambda, \lambda)$.