Chapter 7

Central limit theorems

Recall that a random variable is said to have a **normal distribution** with parameters μ and σ if it has a continuous distribution with density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{for } -\infty < x < \infty.$$

The normal distribution is denoted by $N(\mu, \sigma^2)$. The parameter σ must be positive, otherwise the density would not be positive. The parameter μ can be any real value.

The special case where $\mu = 0$ and $\sigma = 1$ is called the **standard normal**. The density function for this $N(0, 1)$ distribution is usually denoted by the special letter ϕ,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad \text{for } -\infty < x < \infty.$$

For this function to be a well defined density it must integrate to 1, that is,

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi},$$

a result that was derived in Chapter 5.

Using the result from Problem 6.4, you can deduce that X has $N(\mu, \sigma^2)$ distribution if and only if $(X-\mu)/\sigma$ has a standard normal distribution. That is, we can write $X = \mu + \sigma Z$ where Z has a standard normal distribution.

Example <7.1>: The $N(\mu, \sigma^2)$ distribution has expected value μ and variance σ^2.

The normal distribution also has an important stability property: if X and Y are independent, each with a normal distribution, then $X + Y$ also has a normal distribution. This fact will follow from a more general fact about sums of independent random variables

---

Example <7.2>: Suppose X has a continuous distribution with density f and Y has a continuous distribution with density g. If X and Y are independent then the random variable $Z = X + Y$ has a continuous distribution with density

$$h(z) = \int_{-\infty}^{\infty} g(z-x) f(x)\, dx \qquad \text{for all real } z.$$

The integral expression for the density h in terms of f and g is called the **convolution formula**. The next Example shows the formula in action. It also serves as an advertisement for indicator functions. You won't be needing this particular result to understand the general normal approximation. You could safely skip the details.

Example <7.3>: If X and Y are independent, each with the Uniform(0, 1) distribution, find the distribution of $X + Y$.

As promised, the convolution formula also establishes the key fact about sums of independent normals, as recorded in the next Example.

Example <7.4>: If X_1 and X_2 are independent random variables with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

The central limit theorem

The normal approximation to the binomial is just one example of a general phenomenon corresponding to the mathematical result known as the **central limit theorem**. Roughly stated, the theorem asserts:

> *If X can be written as a sum of a large number of relatively small, independent random variables, then it has approximately a $N(\mu, \sigma^2)$ distribution, where $\mu = \mathbb{E}X$ and $\sigma^2 = \text{var}(X)$. Equivalently, the standardized variable $(X - \mu)/\sigma$ has approximately a standard normal distribution.*

See the Appendix for an outline of a proof of a central limit theorem, if you are interested. You can safely ignore the Appendix.

The normal distribution has many agreeable properties that make it easy to work with. Many statistical procedures have been developed under normality assumptions, with occasional offhand references to the central limit theorem to mollify anyone who doubts that all distributions are normal. Modern theory has been much concerned with possible harmful effects of unwarranted assumptions such as normality. The modern fix often substitutes huge amounts of computing for neat, closed-form, analytic expressions; but normality still lurks behind some of the modern data analytic tools.

Example <7.5>: A hidden normal approximation—the boxplot

The normal approximation is heavily used to give an estimate of variability for the results from sampling.

Example <7.6>: Normal approximations for sample means

Things to remember

- If X can be written as a sum of a large number of relatively small, independent random variables, then it has approximately a $N(\mu, \sigma^2)$ distribution, where $\mu = \mathbb{E}X$ and $\sigma^2 = \text{var}(X)$. Equivalently, the standardized variable $(X - \mu)/\sigma$ has approximately a standard normal distribution.

Examples for Chapter 7

<7.1> **Example.** The $N(\mu, \sigma^2)$ is a continuous distribution with density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{for } -\infty < x < \infty.$$

If X has this distribution then

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_{\mu,\sigma}(x)\, dx.$$

Make the change of variable $y = (x - \mu)/\sigma$ to rewrite the integral as

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma y) \exp(-y^2/2)\, \sigma\, dy = \mu \int_{-\infty}^{\infty} \phi(y)\, dy + \sigma \int_{-\infty}^{\infty} y\phi(y)\, dy.$$

We know (from the fact that ϕ is a density function) that the coefficient of μ equals 1. Antisymmetry of $y\phi(y)$ makes it integrate to 0. Thus $\mathbb{E}X = \mu$.

Similarly,

$$\text{var}(X) = \mathbb{E}\,(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_{\mu,\sigma}(x)\,dx = \sigma^2 \int_{-\infty}^{\infty} y^2 \phi(y)\,dy.$$

An integration-by-parts, using the fact that $d\phi(y)/dy = -y\phi(y)$, simplifies the integral,

$$\int_{-\infty}^{\infty} -y\frac{d\phi(y)}{dy}\,dy = \big[-y\phi(y)\big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(y)\,dy = 1.$$

Thus $\text{var}(X) = \sigma^2$.

We could also summarize the calculations by writing $X = \mu + \sigma Z$, with $Z \sim N(0,1)$, then note that

$$\mathbb{E}X = \mu + \sigma\mathbb{E}Z = \mu + \sigma \times 0$$
$$\text{var}(X) = \sigma^2 \text{var}(Z) = \sigma^2 \times 1$$

The changes of variables in the integrals were effectively reducing the problem to the case of a standard normal. □

<7.2> **Example.** Suppose X has a continuous distribution with density f and Y has a continuous distribution with density g. If X and Y are independent show that the random variable $Z = X + Y$ has a continuous distribution with density

$$h(z) = \int_{-\infty}^{\infty} g(z-x) f(x)\,dx \qquad \text{for all real } z.$$

As usual, consider a small, positive δ. Define $F_i = \{i\epsilon \le X < (i+1)\epsilon\}$ for $i = 0, \pm 1, \pm 2, \ldots$, where ϵ is another positive quantity that is much smaller than δ. (More formally, we will be letting ϵ tend to zero while δ stays fixed.) Condition.

$$\mathbb{P}\{z \le X + Y \le z + \delta\} = \sum_{i=-\infty}^{\infty} \mathbb{P}\{z - X \le Y \le z - X + \delta \mid F_i\}\mathbb{P}F_i$$

Conditional on F_i, we know that X is very close to $i\epsilon$. Thus

$$\mathbb{P}\{z - X \le Y \le z - X + \delta \mid F_i\} \approx \mathbb{P}\{z - i\epsilon \le Y \le z - i\epsilon + \delta \mid i\epsilon \le X < (i+1)\epsilon\}$$

On the right-hand side the conditioning now provides no useful information about the behavior of Y: by independence, we can discard the conditioning information. The formula then becomes

$$\mathbb{P}\{z \le X + Y \le z + \delta\} \approx \sum_{i=-\infty}^{\infty} \mathbb{P}\{z - i\epsilon \le Y \le z - i\epsilon + \delta\}\mathbb{P}F_i \approx \sum_{i=-\infty}^{\infty} \delta g(z - i\epsilon)\epsilon f(i\epsilon).$$

Define $H_z(x) = g(z-x)f(x)$. The last sum equals δ times

$$\epsilon \sum_{i=-\infty}^{\infty} H_z(i\epsilon) \approx \int_{-\infty}^{\infty} H_z(x)\,dx.$$

Now let ϵ tend to zero, leaving

$$\mathbb{P}\{z \le X + Y \le z + \delta\} \approx \delta \int_{-\infty}^{\infty} H_z(x)\,dx.$$

Thus $h(z) = \int_{-\infty}^{\infty} H_z(x)\,dx = \int_{-\infty}^{\infty} g(z-x)f(x)\,dx$. □

<7.3> **Example.** If X and Y are independent, each with the Uniform(0, 1) distribution, find the distribution of $X + Y$.

The Uniform(0, 1) has density function $f(x) = \mathbb{I}\{0 < x < 1\}$, that is,

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

The density function h for the distribution of $X + Y$ is given by

$$\begin{aligned} h(z) &= \int_{-\infty}^{\infty} \mathbb{I}\{0 < z - x < 1\}\mathbb{I}\{0 < x < 1\}\, dx \\ &= \int_{-\infty}^{\infty} \mathbb{I}\{0 < x < 1,\ x < z,\ x > z - 1\}\, dx \\ &= \int_{-\infty}^{\infty} \mathbb{I}\{\max(0, z-1) < x < \min(1, z)\}\, dx \end{aligned}$$

If $z \le 0$ or $z \ge 2$ there are no values of x that satisfy the pair of inequalities in the final indicator function; for those cases the indicator function is zero. If $0 < z \le 1$ the indicator becomes $\mathbb{I}\{0 < x < z\}$, so that the corresponding integral equals

$$\int_{-\infty}^{\infty} \mathbb{I}\{0 < x < z\}\, dx = \int_0^z 1\, dx = z.$$

Similarly, if $1 < z < 2$ the integral becomes

$$\int_{-\infty}^{\infty} \mathbb{I}\{z - 1 < x < 1\}\, dx = \int_{z-1}^1 1\, dx = 2 - z.$$

In summary,

$$h(z) = \begin{cases} 0 & \text{if } z \le 0 \text{ or } z \ge 2 \\ z & \text{if } 1 < z \le z \\ 2 - z & \text{if } 1 < z < 2 \end{cases}.$$

More succinctly, $h(z) = \max\big(0, \min(z, 2 - z)\big)$. Maybe it would be better to draw a well labelled picture. □

<7.4> **Example.** If X_1 and X_2 are independent random variables with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Let me simplify the algebra by writing $X_i = \mu_i + \sigma_i Z_i$, where Z_1 and Z_2 are independent standard normals. Then we have $X_1 + X_2 = \mu_1 + \mu_2 + \sigma_1 Z_1 + \sigma_2 Z_2$. It will suffice we show that $W = \sigma_1 Z_1 + \sigma_2 Z_2$ has a $N(0, \sigma_1^2 + \sigma_2^2)$ distribution.

The convolution formula gives the density for the distribution of W,

$$h(z) = \frac{1}{\sigma_1 \sigma_2 2\pi} \int_{-\infty}^{\infty} \exp\left(- \frac{(z-x)^2}{2\sigma_1^2} - \frac{x^2}{2\sigma_2^2} \right) dx.$$

The exponent expands to

$$-\tfrac{1}{2}x^2\left(\sigma_1^{-2} + \sigma_2^{-2}\right) + zx/\sigma_1^2 - \tfrac{1}{2}z^2/\sigma_1^2.$$

Make the change of variable $y = cx$, with

$$c = 1/\sqrt{\sigma_1^{-2} + \sigma_2^{-2}} = \sigma_1\sigma_2/\tau \qquad \text{where } \tau = \sqrt{\sigma_1^2 + \sigma_2^2}.$$

The exponent becomes

$$\begin{aligned} &-\tfrac{1}{2}\left(y^2 - 2zcy/\sigma_1^2 + c^2z^2/\sigma_1^4\right) + \tfrac{1}{2}c^2z^2/\sigma_1^4 - \tfrac{1}{2}z^2/\sigma_1^2 \\ &= -\tfrac{1}{2}\left(y - zc/\sigma_1^2\right)^2 - \tfrac{1}{2}z^2/\tau^2 \end{aligned}$$

The expression for $h(z)$ simplifies to

$$\frac{1}{\tau 2\pi} \exp\left(- \frac{z^2}{2\tau^2} \right) \int_{-\infty}^{\infty} \exp\left(- \tfrac{1}{2}(y - zc/\sigma_1^2)^2 \right) dy.$$

The change of variable $w = y - zc/\sigma_1^2$ then leaves us an integral that equals $\sqrt{2\pi}$.

All the sneaky changes of variable might leave you feeling that the argument is difficult. In fact we didn't have to be so careful. In the original convolution integral we had an exponent of the form $-C_1x^2 + C_2xz - C_3z^2$ for some constants C_1, C_2, C_3. We completed the square to rewrite the exponent as $-C_4(y - C_5z)^2 - C_6z^2$, where y a linear function of x

and C_4, C_5, C_6 were new constants. A change of variable allowed us to integrate out the y, leaving an expression of the form $C_7 \exp(-C_6 z^2)$, which is clearly a $N(0, \tau^2)$ density for some τ. We can calculate τ directly by $\tau^2 = \text{var}(W) = \sigma_1^2 \text{var}(Z_1) + \sigma_2^2 \text{var}(Z_2)$. □

<7.5> **Example.** The boxplot provides a convenient way of summarizing data (such as grades in Statistics 241). The method is:

(i) arrange the data in increasing order

(ii) find the split points

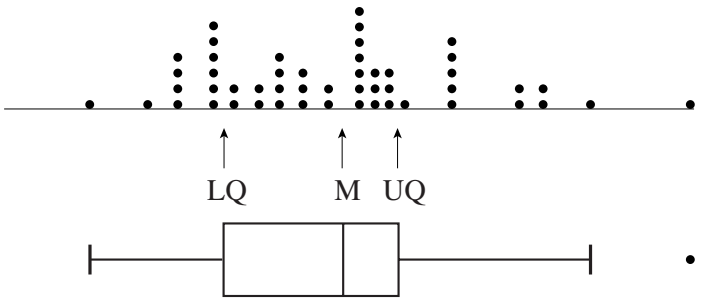
$$\text{LQ} = \text{lower quartile: } 25\% \text{ of the data smaller than LQ}$$
$$\text{M} = \text{median: } 50\% \text{ of the data smaller than M}$$
$$\text{UQ} = \text{upper quartile: } 75\% \text{ of the data smaller than UQ}$$

(iii) calculate IQR (= inter-quartile range) = UQ−LQ

(iv) draw a box with ends at LQ and UQ, and a dot or a line at M

(v) draw whiskers out to UQ + 1.5 × IQR and LQ − 1.5 × IQR, but then trim them back to the most extreme data point in those ranges

(vi) draw dots for each individual data point outside the box and whiskers (There are various ways to deal with cases where the number of observations is not a multiple of four, or where there are ties, or . . .)



Where does the $1.5 \times IQR$ come from? Consider n independent observations from a $N(\mu, \sigma^2)$ distribution. The proportion of observations smaller than any fixed x should be approximately equal to $\mathbb{P}\{W \le x\}$, where W has a $N(\mu, \sigma^2)$ distribution. From normal tables (or a computer),

$$\mathbb{P}\{W \le \mu + .675\sigma\} \approx .75$$
$$\mathbb{P}\{W \le \mu - .675\sigma\} \approx .25$$

and, of course,

$$\mathbb{P}\{W \le \mu\} = .5$$

For the sample we should expect

$$\text{LQ} \approx \mu - .675\sigma$$
$$\text{UQ} \approx \mu + .675\sigma$$
$$\text{M} \approx \mu$$

and consequently,

$$\text{IQR} \approx 1.35\sigma$$

Check that $0.675 + (1.5 \times 1.35) = 2.70$. Before trimming, the whiskers should approximately reach to the ends of the range $\mu \pm 2.70\sigma$. From computer (or tables),

$$\mathbb{P}\{W \le \mu - 2.70\sigma\} = \mathbb{P}\{W \ge \mu + 2.70\sigma\} = .003$$

Only about 0.6% of the sample should be out beyond the whiskers. □

<7.6> **Example.** In Chapter 4 we found the expected value and variance of a sample mean $\bar{Y}$ for a sample of size n from a population $\{y_1, y_2, \ldots, y_N\}$:

$$\mathbb{E}\overline{Y} = \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

and, for sampling with replacement,

$$\operatorname{var}(\overline{Y}) = \frac{\sigma^2}{n} \qquad \text{where } \sigma^2 = \textstyle\sum_{i=1}^{N} (y_i - \overline{y})^2 / N.$$

If Z has a $N(0, 1)$ distribution,

$$\mathbb{P}\{-1.96 \le Z \le 1.96\} \approx 0.95.$$

The standardized random variable $(\overline{Y}-\overline{y})/\sqrt{\sigma^2/n}$ is well approximated by the $N(0, 1)$. Thus

$$\mathbb{P}\left\{-\frac{1.96\sigma}{\sqrt{n}} \le \overline{Y} - \overline{y} \le \frac{1.96\sigma}{\sqrt{n}}\right\} \approx 0.95.$$

Before we sample, we can assert that we have about a 95% chance of getting a value of $\overline{Y}$ in the range $\overline{y} \pm 1.96\sigma/\sqrt{n}$. (For the post-sampling interpretation of the approximation, you should take Statistics 242/542.)

Of course, we would not know the value σ, so it must be estimated.

For sampling without replacement, the variance of the sample mean is multiplied by the correction factor $(N - n)/(N - 1)$. The sample mean is no longer an average of many *independent* summands, but the normal approximation can still be used. (The explanation would take us beyond 241/541.) □

Appendix: Lindeberg's method for the Central Limit Theorem

We have $X = X_1 + X_2 + \ldots + X_n$, a sum of a lot of small, independent contributions. If all the X_i's are normally distributed, repeated appeals to Example <7.4> show that X is also normally distributed.

If the X_i's are not normal, we replace them one at a time by new independent random variables Y_i for which $\mathbb{E}Y_i = \mathbb{E}X_i$ and $\operatorname{var}(Y_i) = \operatorname{var}(X_i)$. It is easy to use Taylor's theorem to track the effect of the replacement if we consider smoooth functions of the sum.

For example, suppose h has a lot of bounded, continuous derivatives. Write S for $X_1 + \ldots + X_{n-1}$. Then

$$\begin{aligned}\mathbb{E}h(X) = \mathbb{E}h(S + X_n) &= \mathbb{E}\left(h(S) + X_n h'(S) + \tfrac{1}{2}X_n^2 h''(S) + \tfrac{1}{6}X_n^3 h'''(S) + \ldots\right)\\ &= \mathbb{E}h(S) + \mathbb{E}X_n \mathbb{E}h'(S) + \tfrac{1}{2}\mathbb{E}(X_n^2)\mathbb{E}h''(S) + \tfrac{1}{6}\mathbb{E}(X_n^3)\mathbb{E}(h'''(S)) + \ldots\end{aligned}$$

In the last line, I have used the independence to factorize a bunch of products.

Exactly the same idea works for $h(S + Y_n)$. That is,

$$\mathbb{E}h(S + Y_n) = \mathbb{E}h(S) + \mathbb{E}Y_n \mathbb{E}h'(S) + \tfrac{1}{2}\mathbb{E}(Y_n^2)\mathbb{E}h''(S) + \tfrac{1}{6}\mathbb{E}(Y_n^3)\mathbb{E}(h'''(S)) + \ldots$$

Subtract the two expansions, noting the cancellations caused by the matching of first and second moments for X_n and Y_n.

$$\mathbb{E}h(S + X_n) - \mathbb{E}h(S + Y_n) = \tfrac{1}{6}\mathbb{E}(X_n^3)\mathbb{E}(h'''(S)) + \ldots - \tfrac{1}{6}\mathbb{E}(Y_n^3)\mathbb{E}(h'''(S)) + \ldots$$

A similar argument works if we replace the X_{n-1} in $\mathbb{E}h(S + Y_n)$ by its companion Y_{n-1}. And so on. After we swap out all the X_i's we are left with

$$Eh(X) - \mathbb{E}h(Y_1 + Y_2 + \ldots Y_n) = \text{ a sum of quantities of third, or higher order.}$$

A formal theorem would give a precise meaning to how small the X_i's have to be in order to make the "sum of quantities of third, or higher order" small enough to ignore.

If you were interested in expectations $\mathbb{E}h(X)$ for functions that are not smooth, as happens with $\mathbb{P}\{X \le x\}$, you would need to approximate the non-smooth h by a smooth function for which Lindeberg's method can be applied.