The multinomial distribution

Consider a *generalised Bernoulli experiment* with k possibilities $\{b_1, b_2, \ldots, b_k\}$ where every trial can result in one of them.

Let b_i occur with probability $\theta_i > 0$ and $\sum_{i=1}^{k} \theta_i = 1$.

Consider n independent trials and let
X_1 be the rv that counts the number of b_1's,
X_2 the number of b_2's, $\ldots$
X_k the number of b_k.

The probability of observing n_1 b_1's, n_2 b_2's, etc., n_k b_k's is

$$P(X_1 = n_1, \ldots, X_k = n_k) = \tfrac{n!}{n_1! \cdots n_k!} \theta_1^{n_1} \cdots \theta_k^{n_k},$$

and the rve $(X_1, \ldots, X_k)$ is said to follow a multinomial distribution, denoted $(X_1, \ldots, X_k) \sim \mathrm{Mn}(n; \theta_1, \ldots, \theta_k)$.

The multinomial distribution

Example 7 (Blood type)

In the human population, 48% have type O, 38% have type A, 10% have type B and 4% have type AB. In a sample of 20 people, what is the probability that 7 have type O, 7 have type A, 4 have type B and 2 have type AB?

The probability is

$$\frac{20!}{(7!)^2 4! 2!} 0.48^7 0.38^7 0.1^4 0.04^2 = 0.00214.$$

The multinomial distribution

(i) with $k = 2$ $\mathrm{Mn}(n; \theta_1, \theta_2) = \mathrm{Bin}(n, \theta)$;

(ii) $X_j \sim \mathrm{Bin}(n, \theta_i)$;

(iii) Every d-subvector $(X_{i_1}, \ldots, X_{i_d})$ of X, $d \leq k$ is multinomial;

(iv) If also $Y \sim \mathrm{Mn}(n_y; \theta_1, \ldots, \theta_k)$ then $Y + X \sim \mathrm{Mn}(n_z; \theta_1, \ldots, \theta_k)$, with $n_z = n + n_y$.

(v) If $X_1, \ldots, X_k$ are independent and $X_i \sim \mathrm{Poi}(\lambda_i)$, then $X_1, \ldots, X_k$ given their sum, is multinomial:

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \,\middle|\, \sum_{i=1}^{k} X_i = n\right) = \mathrm{Mn}(n; \theta_1, \ldots, \theta_k),$$

where $\theta_i = \lambda_i / \sum_{j=1}^{k} \lambda_j$, $i = 1, \ldots, k$.

(vi) $E(X) = (n\theta_1, \ldots, n\theta_k)$ and for all $i, j = 1, \ldots, k$

$$\mathrm{cov}(X_i, X_j) = \begin{cases} -n\theta_i\theta_j & \text{if } i \neq j \\ n\theta_i(1 - \theta_j) & \text{if } i = j. \end{cases}$$

Probability inequalities

Markov's inequality: If X is a non-negative rv s.t. $E(X) < \infty$, then

$$P(X > t) \leq \frac{E(X)}{t}, \quad \text{for any } t > 0.$$

Chebyshev's inequality: If X is a non-negative rv s.t. $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$, are both finite, then

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}, \quad \text{for any } t > 0.$$

Convergence of rv's

Given $X_1, X_2, \ldots$, a sequence of rv's we wish to say something about its **limiting behaviour**.

This is important, since statistics and data mining are all about gathering data, and we will often be interested in what happens as we gather more and more data.

You may recall the definition of the convergence of a sequence or numbers: A sequence $x_1, x_2, \ldots$, is convergent to a number x if for any $\epsilon > 0$, there is $N \in \mathbb{N}$ s.t. $|x_n - x| < \epsilon$ for all $n \geq N$.

When dealing with rv's, the concept of convergence becomes more subtle.

For example, if $x_1, x_2, \ldots$, with $x_i = x$ for all i, then $\lim x_n = x$. On the other hand, if $X_1, X_2, \ldots$ with $X_i \sim \mathrm{N}(0, 1)$, all independent, then we may think that the sequence 'converges' to $X \sim \mathrm{N}(0, 1)$; but this is not right as $P(X_n = X) = 0$, for all $n \ldots$

Motivating example

Consider $X_1, \ldots, X_n$ fully independent, with each $X_i \sim \text{Unif}(0, 1)$, and let $\overline{Y}_n = (X_1 + \cdots + X_n)/n$.

We are interested in the distribution of Y_n as n gets large.

Although we could use probability calculus, we can get easily get a good idea via simulation. Let's switch to `R`...

Types of convergence

Let $X_1, \ldots, X_n$ be a sequence of rv's, with $X_i \sim F_n$ and let the rv $X \sim F$. Then

X_n converges to X in probability, written $X_n \xrightarrow{P} X$, if

$$\lim_{n\to\cancel{0}} P(|X_n - X| > \epsilon) = 0, \quad \text{for all} \quad \epsilon > 0.$$

∞

X_n converges to X in distribution, $X_n \xrightarrow{d} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t), \quad \text{at all } t \text{ where } F \text{ is continuous.}$$

Note: the limiting rv X can also be a point mass distribution, i.e. $P(X = \mu) = 1$ for some μ.

Example 8

Let $X_n \sim \mathrm{N}(0, 1/n)$. All terms have mean zero and variance $\to 0$, so the sequence must converge at zero. Let's check.

Consider convergence in probability first. For any $\epsilon > 0$, using Chebyshev's inequality

$$0 \le P(|X_n| > \epsilon) = P(|X_n| \ge \epsilon) \le \frac{1/n}{\epsilon^2}.$$

Now $\lim_{n\to\infty} \frac{1/n}{\epsilon^2} = 0$, thus by the properties of limits, $P(|X_n| > \epsilon) \to 0$ as $n \to \infty$.

Example 8 (cont'd)

And now convergence in distribution.
Let Z denote the standard normal rv and F be the df of a point mass at zero. We have $\sqrt{n}X_n \sim \mathrm{N}(0,1)$, thus

$$F_n(t) = P(X_n \leq t) = P(\sqrt{n}X_n \leq \sqrt{n}t) = P(Z \leq \sqrt{n}t).$$

So,

$$\lim_{n\to\infty} F_n(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{otherwise.} \end{cases}$$

Thus, except for $t = 0$, we see that $F_n(t) \to F(t)$. We don't have to worry for $t = 0$, since at this point $F(t)$ has a jump, thus we conclude that $X_n \xrightarrow{d} 0$.

Algebra of sequences of rv

For $X_1, \dots, X_n$ a sequence of rv's with each term having df F_n, and let X be a rv with df F. Then

(i) If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$.

(i) If $X_n \xrightarrow{d} X$ and if $P(X = c) = 1$ for some real c, then $X_n \xrightarrow{P} X$.

Algebra of sequences of rv (cont'd)

Let X_n, X, Y_n, Y, be rv's and g a continuous function. Then

(i) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$.

(ii) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ then $X_n + Y_n \xrightarrow{d} X + c$.

(iii) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n Y_n \xrightarrow{P} XY$.

(iv) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ then $X_n Y_n \xrightarrow{d} cX$.

(v) If $X_n \xrightarrow{P} X$ then $g(X_n) \xrightarrow{P} g(X)$.

(vi) If $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$.

Parts (ii) and (iv) are known as Slutzky's lemma.

The Law of Large Numbers

Also called Weak LLN states that if $X_1, \ldots, X_n$ are iid rv with $E(X_1) = \mu$, $\sigma^2 = \text{var}(X_1)$, then

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} \mu.$$

Note that $E(\overline{X}_n) = E((X_1 + \cdots + X_n)/n) = \mu$. Thus, the WLLN says that the distribution of the average of the random variables collapses to the population average μ as n goes to infinity.

The Law of Large Numbers

Example 9

Consider flipping a coin with probability of heads p and let X_i denote the outcome of the ith toss; this can take on only 0 or 1. Thus $p = P(X_i = 1) = E(X_i)$.

According to WLLN, $\overline{X}_n \xrightarrow{P} p$, this means that the distribution of $\overline{X}_n$ is more and more concentrated around p as n diverges.

Suppose that $p = 1/2$. How large n should be s.t. $P(.4 \leq \overline{X}_n \leq .6) \geq .7$? $E(\overline{X}_n) = 1/2$, furthermore, $\mathrm{var}(\overline{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$. Using Chebyshev's inequality we have...

Example 9 (cont'd)

$$P(.4 \leq \overline{X} \leq .6) = P(|\overline{X}_n - 1/2| \leq .1) = 1 - P(|\overline{X}_n - 1/2| > .1)$$
$$\geq 1 - \frac{1}{4n(.1)^2} = 1 - \frac{25}{n},$$

and the last inequality is satisfied for $n \geq 84$.

The Central Limit Theorem

WLLN tells us only **where** will the distribution of $\overline{X}_n$ eventually **collapse to**: μ; no more, no less.

The CLT tells us **what** is the **shape** of this distribution as n diverges.

CLT: Let $X_1, \ldots, X_n$ be independent rv's with $\mu = E(X_i)$, $\sigma^2 = \mathrm{var}(X_i)$. Then

$$Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\mathrm{var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z,$$

where $Z \sim \mathrm{N}(0, 1)$. Another way to say this is

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x.$$

So, we can **approximate** probability statements about $\overline{X}_n$ **by** probability statements about Z.

The Central Limit Theorem[2]

Example 10

Suppose that the number of bugs in a computer program has a Poisson distribution with mean 5. Given 125 programs, what is the probability that the average number of bugs is less than 5.5?

We can approximate this by CLT. Now $\mu = E(X_1) = 5$, $\mathrm{var}(X_1) = 5$, thus

$$\begin{aligned} P(\overline{X}_n < 5.5) &= P = \left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma} \right) \\ &\doteq P(Z \leq 2.5) = .9938. \end{aligned}$$

The symbol "$\doteq$" stands for "asymptotically, i.e. $n \to \infty$, equal to".

[2] There is also a multivariate version of the CLT, in which the limiting distribution is the standard multivariate normal.

The Delta Method

Useful when we know X_n has a limiting normal distribution and we wish to find the limiting distribution of $g(Y_n)$, where g is any smooth function.

The Delta Method: Let X_n be a sequence of rv s.t.

$$\frac{\sqrt{n}(X_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1),$$

and that g is differentiable s.t. $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} \mathrm{N}(0,1).$$

In other terms,

$$X_n \dot{\sim} \mathrm{N}(\mu, \sigma^2/n) \quad \Rightarrow \quad g(X_n) \dot{\sim} \mathrm{N}\left(g(\mu), (g'(\mu))^2\sigma^2/n\right),$$

"$\dot{\sim}$" stands for "asymptotically distributed as".

The Multivariate Delta Method

Serves the same purpose as DM in the case of r.ve.'s. In particular,

The Multivariate Delta Method: Let $X_n = (X_{n1}, X_{n2}, \ldots, X_{np})$, $n = 1, 2, \ldots$, be a sequence p-dimensional random vectors such that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathrm{N}_p(0, \Sigma).$$

Suppose that $g(x) = (g_1(x), \ldots, g_k(x))$, with $g : \mathbb{R}^p \to \mathbb{R}^k$, $k \leq p$, is such that the $k \times p$ matrix of partial derivatives

$$B(x) = [b_{ij}] = \left[\frac{\partial g_i(x)}{\partial x_j}\right], \quad x = (x_1, \ldots, x_p),$$

are continuous and do not vanish in a neighbour of μ and let $B_\mu = B(\mu)$. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathrm{N}_{\underset{k}{\cancel{p}}}(0, B_\mu \Sigma B_\mu^{\mathrm{T}}).$$

Delta Method

Example 11

Let $X_1, \dots, X_n$ be independent rv's with mean μ and variance σ^2. By CLT,

$$\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{d} \mathrm{N}(0, 1).$$

Let $Y_n = g(\overline{X}_n) = \exp(\overline{X}_n)$. Then $g'(\mu) = e^{\mu}$ and the delta method implies that

$$Y_n \dot{\sim} \mathrm{N}(e^{\mu}, e^{2\mu}\sigma^2/n).$$

Multivariate Delta Method

Example 12

Let $X_1, \dots, X_n$ be independent rve with $X_n = (X_{1n}, X_{2n})$ having mean $\mu = (\mu_1, \mu_2)$ and variance Σ. Let

$$\overline{X}_1 = \tfrac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \overline{X}_2 = \tfrac{1}{n}\sum_{i=1}^{n} X_{2i},$$

and define $Y_n = \overline{X}_1\overline{X}_2$. Thus $Y_n = g(\overline{X}_1, \overline{X}_2)$, with $g(x_1, x_2) = x_1x_2$. By the (multivariate) CLT,

$$\sqrt{n}\left(\begin{array}{c} \overline{X}_1 - \mu_1 \\ \overline{X}_2 - \mu_2 \end{array}\right) \xrightarrow{d} \mathrm{N}_2(0, \Sigma).$$

Now $B(x) = (x_2, x_1)$ and thus $B_\mu = (\mu_2, \mu_1)$ and

$$B_\mu \Sigma B_\mu^{\mathrm{T}} = \mu_2\sigma_1^2 + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_2^2,$$

therefore

$$\sqrt{n}(\overline{X}_1\overline{X}_2 - \mu_1\mu_2) \dot{\sim} \mathrm{N}(0, \mu_2\sigma_1^2 + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_2^2).$$