

Example (inverse binomial sampling). Consider sampling from a population of bats until the fifth animal with rabies is obtained. The rabies rate is supposed to be 0.01. We sampled 98 to get 5. Should we be concerned?

Solution: consider the r.v. $X_k = \text{"\# bats sampled"}$ ($k = 5$).

Natural estimator for rabies rate p : $\hat{p} = 5/98 = k/X_k \approx 0.05$.

Is it likely to observe the rate 0.05 if in fact $p = 0.01$? Is $P_{0.01}(X_5 \leq 98)$ large or small?

$N_k = X_k - k \sim$ negative binomial(k, p) with $k = 5, p = 0.01$.

$P(X_5 \leq 98) = P(N_5 + 5 \leq 98) = P(B_{98} \geq 5)$ where $B_{98} \sim \text{bin}(98, 0.01)$.

We can use the Poisson approximation to obtain $P(X_5 \leq 98) = 1 - P(B_{98} \leq 4) \approx 0.003$. *Conclusion:* a sample size of 98 or smaller is very unlikely if $p = 0.01$; this suggests $p > 0.01$.

Remark. Let $N_n \sim$ negative binomial($k, \lambda/n$) with k fixed. Then, as $n \rightarrow \infty$,

$$\frac{N_n}{n} \xrightarrow{\mathcal{D}} X \sim \text{gamma}(\alpha, \beta) \quad \text{with } \alpha = k, \beta = \frac{1}{\lambda}$$

(“ $\xrightarrow{\mathcal{D}}$ ” means convergence in distribution), i.e. $P(N_n \leq nx) \rightarrow F_X(x)$ ($n \rightarrow \infty$). This can, for example, be shown by using the mgf. [The pdf of the gamma(α, β) distribution ($\alpha > 0, \beta > 0$): $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} 1_{(0, \infty)}(x)$ with $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.]

Theorem. Let $X_k \sim \text{gamma}(k, 1/\lambda)$, $Y_t \sim \text{Poisson}(\lambda t)$. Then $P(X \leq t) = P(Y_t \geq k)$.

Idea of proof: apply the previous theorem (relationship negative binomial and binomial) and take the limit.

Poisson process

Divide the time axis into equally sized units; divide each time unit into n equally sized subintervals:

$$0 - \dots - 1 \underbrace{\quad \quad \quad}_{n \text{ subintervals}} 2 - \dots - 3 - \dots - \quad \quad \quad (\text{time units})$$

Consider independent trials (“occurrence yes/no” in each subinterval). Assume we expect λ successes for each time unit; then $p = P(\text{"success in a trial"}) = \lambda/n$. Suppose the subintervals are small and $P(\text{"2 or more successes per trial"}) \approx 0$. Then we have: time until k -th success $\approx \frac{1}{n} \cdot \#$ trials until k -th successes $\approx \frac{1}{n} \cdot$

N_n with $N_n \sim$ negative binomial(k, p) and, as shown earlier, $N_n/n \xrightarrow{\mathcal{D}} X \sim \text{gamma}(k, 1/\lambda)$. Therefore,

$$\text{time until } k\text{-th success} \sim \text{gamma}(k, 1/\lambda)$$

Now consider a fixed interval $[0, t]$ where t is a real number ($\approx nt$ trials). Then

$$X_t = \text{"\# successes in the interval } [0, t]\text{"} \approx \text{bin}(nt, p) \xrightarrow{\mathcal{D}} \text{Poisson}(t\lambda) \quad (\lambda = np).$$

$(X_t)_t$ is the **Poisson process**.

Note that non-overlapping intervals $[0, t]$, $[t, t + s]$ have independent numbers of occurrence since the trials are independent.

Definition. Consider a Poisson process with occurrence rate λ per time unit.

(i) The distribution of the **number of occurrences** in the interval $[0, t]$ is the **Poisson**(λt) distribution.

(ii) The distribution of the **waiting time** until the k -th occurrence is the **gamma** distribution with parameters k and $1/\lambda$.

Special case $k = 1$: **exponential** distribution.

3.2 Random sampling

Definition. Suppose a population of size N has $M = pN$ “positives” (successes, 1’s) and $(1 - p)N$ “negatives” (failures, 0’s). Consider a sample of n individuals.

The **binomial distribution** ($\text{bin}(n, p)$) is the distribution of the number of positive responses in a simple random sample *with replacement*.

If we sample *without replacement*, the distribution is **hypergeometric**(n, N, M).

Example.

From the quality control example (SWOR) with “ $X = \#$ defectives” we know $P(X = x) = \binom{M}{x} \binom{N-M}{n-x} / \binom{N}{n}$. We have $\binom{N}{n}$ equally likely *unordered* samples.

With order identified: $|\mathcal{S}| = \binom{N}{n} n!$ (equally likely samples). Let $A_i =$ “ i -th selection is positive”. Then, for $s \in \mathcal{S}$,

$$Y(s) = \sum_{i=1}^n 1_{A_i}(s) \quad \text{is the hypergeometric variable.}$$

We will now compute EY without using the pmf.

$$\begin{aligned} EY &= \sum_{s \in \mathcal{S}} Y(s) \cdot \frac{1}{|\mathcal{S}|} = \frac{1}{n! \binom{N}{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^n 1_{A_i}(s) = \sum_{i=1}^n P(A_i), \\ P(A_i) &= \frac{1}{n! \binom{N}{n}} \sum_{s \in \mathcal{S}} 1_{A_i}(s) = \frac{1}{n! \binom{N}{n}} M(n-1)! \binom{N-1}{n-1} \\ &= \frac{1}{n! \frac{N!}{(N-n)!n!}} M \frac{(N-1)!}{(N-n)!} = \frac{M}{N} = p. \end{aligned}$$

This gives $EY = \sum_{i=1}^n p = np$. Similarly (calculate EY^2 first), $\text{Var}(Y) = \frac{N-n}{N-1} np(1-p)$, which is smaller than the binomial variance since $(N-n)/(N-1) \leq 1$ ($\Leftrightarrow n \geq 1$).

Note. There is almost no difference between SWR and SWOR for large populations (large N). If Y denotes the number of successes, then $E(Y) = np$ for both sampling schemes and $\text{Var}(Y)$ is almost the same since $(N-n)/(N-1) = N/(N-1) - n/(N-1) \approx 1 - 0 = 1$.

Theorem (proof exercise). Consider $Y_N \sim \text{hypergeometric}(n, N, pN)$ and $Y \sim \text{bin}(n, p)$. Then $Y_N \xrightarrow{D} Y$ as $N \rightarrow \infty$. In particular, $P(Y_N = y) \rightarrow P(Y = y)$ for every $y = 0, 1, 2, \dots$