Applied Statistics

Sofia Triantafillou

sof.triantafillou@gmail.com

University of Crete
Department of Mathematics and Applied Mathematics

Lecture Summary

- Useful Families of Distributions (all in Chapter 5):
  - Discrete: Bernoulli, Binomial, Geometric.
  - Poisson Distribution
  - Exponential Distribution.
  - Normal Distribution.
- Recap: Central Limit Theorem.

Families of Distributions

- Probability function notation: $f(x|\text{parameters})$.
- Parameter space.
- Mean, variance.
- Types of experiments.

Bernoulli distributions

A r.v. X has the Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The probability function (pf) of X is

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & otherwise \end{cases}$$

- An experiment with two outcomes: "success", "failure", X = number of successes.
- Parameter space: $p \in [0, 1]$.
- $E(X) = p, Var(X) = p(1 - p)$.

Binomial distributions

A r.v. X has the Binomial distribution with parameters n and p a if the probability function (pf) of X is

$$f(x|p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \textit{otherwise} \end{cases}$$

- n repetitions of an experiment with two outcomes: "success", "failure", X = number of successes.
- Parameter space: n positive integer, $p \in [0, 1]$.
- $E(X) = \quad , Var(X) = \quad .$

Geometric distributions

A r.v. X has the Geometric distribution with parameters n and p a if the probability function (pf) of X is

$$f(x|p) = \begin{cases} f(x|p) = p(1-p)^x & x = 0, 1, \ldots n \\ 0 & otherwise \end{cases}$$

- An experiment with two outcomes: "success", "failure", X = number of failures before the first success.
- Parameter space $p \in [0, 1]$.
- $E(X) = \frac{1-p}{p}, Var(X) = \frac{1-p}{p^2}$.

Geometric distributions

Geometric distributions are memoryless:

Theorem

Let X have the geometric distribution with parameter p, and let $k \geq 0$. Then for every integer $t \geq 0$,

$$P(X = k + t | X \geq k) = P(X = t).$$

The Poisson distributions

Let $\lambda > 0$. A random variable X follows the *Poisson distribution* with mean λ if the p.m.f. of X is as follows:

$$f(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & otherwise. \end{cases}$$

- Parameter space: $\lambda \in [0, \infty)$.
- $E(X) = p, Var(X) = p(1 - p)$

The Poisson Distribution

The Poisson distribution is useful for modeling uncertainty in counts / arrivals.

Examples:

- How many calls arrive at a switch board in one hour?
- How many busses pass while you wait at the bus stop for 10 min?
- How many customers will enter a store in 15 minutes?

Properties of the Poisson

Theorem (Sum of Poissons is a Poisson.)

Let $X_1, \ldots X_k$ are independent and if X_i has the Poisson distribution with mean $\lambda_i (i = 1, \ldots, k)$, then the sum $X_1 + \cdots + X_k$ has the Poisson distribution with mean $\lambda_1 + \cdots + \lambda_k$.

Theorem (Approximation to the Binomial)

For each integer n and each $0 < p < 1$, let $f(x|n, p)$ denote the pf of the Binomial distribution with parameters n and p, and let $f(x|\lambda)$ denote the pf of the Poisson distribution with mean λ. Let $\{p_n\}_1^\infty$ be a sequence of numbers between 0 and 1 such that $lim_{n\to} = \lambda$. Then

$$lim_{n\to\infty} f_{X_n}(x|n, p_n) = f(x|\lambda)$$

When the value of n is large, and the value of p is very small, the Poisson with mean np is a good approximation for the Binomial with parameters n and p.

The Exponential Distributions

Let $\beta > 0$. A random variable X follows the *exponential distribution* with parameter β if it has a continuous distribution with pf.:

$$f(x|\beta) = \begin{cases} \beta e^{-\beta x} & x > 0, \\ 0 & \textit{otherwise}. \end{cases}$$

- Parameter space: $\beta \in [0, \infty)$.
- Find $E(X)$, $Var(X)$.

Normal Distributions

- Standard normal: $\mathcal{N}(0,1) : f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- Normal with mean μ and variance σ^2:
  $\mathcal{N}(\mu,\sigma^2) : f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- $E(X) = \mu,\ Var(X) = \sigma^2$.

Theorem (Linear transformations of a normal are normal)

If $X \sim \mathcal{N}(\mu,\sigma^2)$, then $\alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$

Theorem (The sum of independent normals is normal)

If the random variables $X_1, \ldots, X_k$ are independent and if $X_i \sim N(\mu_i, \sigma_i^2)$ then $X_1 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2)$

Calculating probabilities with the Normal Distribution

- We want to estimate $P(X \leq a)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$
- No closed form for $\int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dt$
- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$
- $P(X \leq a) = P(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$.

Central Limit Theorem

Theorem (Central Limit Theorem)

If the random variables $X_1, \ldots, X_n$ form a random sample of size n from a given distribution with mean μ and variance σ^2 ($0 < \sigma^2 < \infty$), then for each fixed number x

$$lim_{n\to\infty} P(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}) = \Phi(x),$$

where Φ denotes the c.d.f. of the standard normal distribution.

In practice, it often holds for small n.

Frame Title

- We need to test 1000 people for a rare disease (affects two in 1000 people).
- Each test requires a small amount of blood, and it is guaranteed to detect the disease if it is anywhere in the blood.
- Strategy 1: Test 1000 people
- Strategy 2: Split in 10 groups of 100, combine their bloods and test them. (10 tests). If any of them tests positive, test all people in that group.
- What is the expected number of tests for Strategy 2?