Daniel **Fitz**
(43961229)

University of Queensland

STAT2203 – Probability and Statistics for Engineering

STAT2203 Lecture Notes

Table of Contents

Sums and Extremes of Independent Random Variables 2
Multivariate Normal Distribution 2
Affine Combinations of Normals Example 2
Sums of Independent Random Variables 2
Example 2
Law of Large Numbers 2
Central Limit Theorem 3
Approximating Binomial by Normal 3
Example 3
Approximating via the CLT 4
Extremes of Independent Random Variables 4
Summary 4
Statistics, Likelihood, and Estimation 4
Sums and Extremes of Independent Random Variables 4
Statistics 5
Likelihood 5
Likelihood Example 5
Maximum Likelihood Estimation 5
ML Estimation Example: Binomial Probability 6
Summary 6

List of Tables

Placeholder for table of contents 0

List of Figures

Figure 1: Approximating Binomial by Normal 3

Sums and Extremes of Independent Random Variables

Multivariate Normal Distribution

- Jointly Gaussian Random Variables; as affine transform of vector of independent standard normals, examples
- Expectation of Vector- and Matrix-valued RVs; application to multivariate normal
- Affine combinations of independent normals; result, examples

Affine Combinations of Normals Example

Exercise: Let $X_1, ..., X_n$ ~ $N(u,o^2)$ represent repeated measurements. Find is the distribution of the average measurement

$$Y = \frac{X_1 \cdots + X_n}{n}$$

Sums of Independent Random Variables

Law of Large Numbers and the **Central Limit Theorem**. Both theorems deal with **Sums of Independent Random Variables**. They arise for example in the following situations:

1) We flip a (biased) coin infinitely many times. Let $X_i = 1$ if the ith flip is "heads" and $X_i = 0$ otherwise. In general we do not know $p = P(X_i = 1)$. However, using the outcomes $x_1, ..., x_n$, we could estimate p by $(x_1 + ... + x_n)/n$
2) A certain machine needs to work continuously. The machine has one component that is very unreliable. This component is replaced immediately upon failure. Suppose there are n such (spare) components. If we denote the component lifetimes by $X_1, ..., X_n$, then the lifetime of the machine is given by $X_1 + ... + X_n$.
3) We weigh 20 randomly selected people. The average weight of the group is $(X_1 + ... + X_{20})/20$

Let $X_1, ..., X_n$ be independent and identically distributed random variables. For each n let

$$Sn = X1 \cdots + Xn$$

Let $EX_i = u$ and $Var(X_i) = o^2$ (assuming that these are finite).

Some easy results are:

$$\mathbb{E}S_n = n\mathbb{E}X_1 = n\mu$$

and, by the independence of the summands,

$$Var(S_n) = n\ Var(X_1) = n\sigma^2$$

If we know the pdf or pmf of X_i, then we can (in principle) determine the pdf or pmf of S_n. The easiest way is to use **transform** techniques (Laplace transform, Characteristic function, etc).
An important property of these transforms is that the transform of the **sum** of independent random variables is equal to the **product** of the individual transforms.

Example

Example: Suppose each X_i ~ $Exp(lambda)$. The Laplace transform of X_i, say L is given by

$$L(s) = \mathbb{E}e^{-sX_i} = \frac{\lambda}{\lambda + s}$$

The Laplace transform of S_n, is given by

$$\begin{aligned}\mathbb{E}e^{-sS_n} &= \mathbb{E}e^{-s(X_1+\cdots+X_n)} \\ &= \mathbb{E}e^{-sX_1}\cdots\mathbb{E}e^{-sX_n} = (L(s))^n \\ &= \left(\frac{\lambda}{\lambda+s}\right)^n\end{aligned}$$

Using the uniqueness of Laplace transforms, this shows that S_n has a Gamma(n, lambda) distribution (Erlang distribution)

Law of Large Numbers

Consider the coin flip example. We expect that S_n/n is close to the unknown p for large n.
We know this happens "empirically".

In general, we expect S_n/n to be close to u. Does this happen in our mathematical model?
By *Chebyshev's inequality* we have for all $e > 0$,

$$\mathbb{P}\left(\mid \frac{S_n}{n} - \mu \mid > \epsilon \right) \leq \frac{Var(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to$ *infinite*.

In other words the probability that S_n/n is more than *e* away from *u* can be made arbitrarily small by choosing *n* large enough.
This is the **Weak Law of Large Numbers**.

There is also a **Strong Law of Large Numbers:**

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{S_n}{n} = \mu \right) = 1$$

as $n \to$ *infinite*

Central Limit Theorem

The Central Limit Theorem states, roughly, this: The **sum** of a large number of **iid** random variables has **approximately** a **Gaussian** distribution.
More precisely, it states that for all *x*

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x \right) = \Phi(x)$$

where *Phi* is the cdf of the standard normal distribution.

Approximating Binomial by Normal

Using the CLT we thus find the following important approximation:
Let *X ~ Bin(n, p)*. For large *n*, we have

$$\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k)$$

where *Y ~ N(np, np(1 - p))*.

As a rule of thumb, the approximation is accurate if both *np* and *n(1 - p)* are larger than 5.

We can improve on this somewhat by using a continuity correction, as illustrated by the following graph for the pmf of the *Bin(10, 1/2)* distribution.

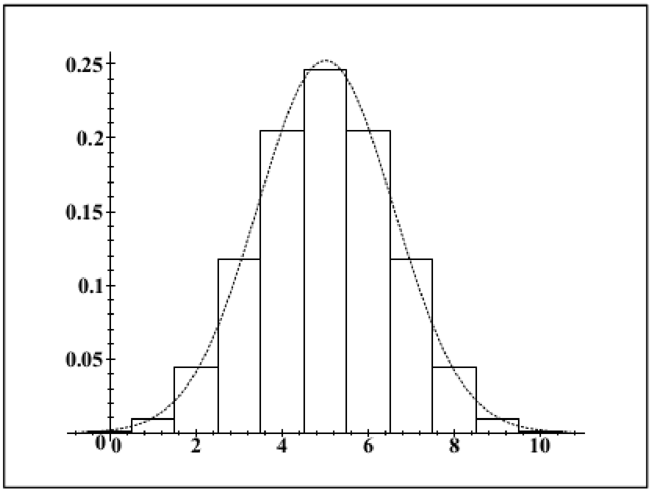

Figure 1: Approximating Binomial by Normal

For example,

$$\mathbb{P}(X = k) \approx \mathbb{P}(k - \frac{1}{2} \leq Y \leq k + \frac{1}{2})$$

Example

Example: Let *X ~ Bin(200, 0.51)*, and suppose we wish to to calculate *P(X <= 99)*.
Let *Y ~ N(200 x 0.51, 200 x 0.51 x 0.49)*, and let *Z* be standard normal. Using the CLT we have

$$\begin{aligned}\mathbb{P}(X \leq 99) &\approx \mathbb{P}(Y \leq 99)\\ &= \mathbb{P}\left(\frac{Y-102}{\sqrt{49.98}} \leq \frac{99-102}{\sqrt{49.98}}\right)\\ &= \mathbb{P}(Z \leq -0.4243) = 1 - \mathbb{P}(Z \leq 0.4243)\\ &= 0.3357\end{aligned}$$

Using the continuity correction we find

$$\mathbb{P}(X \leq 99) \approx \mathbb{P}(Y \leq 99 + \frac{1}{2}) = 0.3618$$

Approximating via the CLT

Exercise: The number of calls X arriving at a call centre during an hour has a *Poi(100)* distribution.
Show, using probability generating functions, that X has the same distribution as $X_1 + ... + X_{100}$, where $X_1, ..., X_{100}$ are independent *Poi(1)*-distributed random variables.
Use this fact to approximate (with the CLT) the probability that there are more than 130 arrivals during an hour

Extremes of Independent Random Variables

In addition to the average behaviour of iid variates $X_1, ..., x_n$, we are often interested in the extremes – that is, how the largest (or smallest) variate behaves.

If $M = max\{X_1, ..., X_n\}$, we have seen (by example) that

$$\begin{aligned}F_M(m) &= \mathbb{P}(M \leq m) = \mathbb{P}(X_1 \leq m, \dots, X_n \leq m)\\ &= \mathbb{P}(X_1 \leq m)^n = (F_X(m))^n\end{aligned}$$

What distribution does M have, as $n \to$ *infinite*

Remark: It turns out that, when M is suitably shifted and scaled, there are essentially three possibilities (listed here for completeness). The Gumbel distribution (*u element R, o > 0*):

$$f(x) = \frac{1}{\sigma} exp\left[-\frac{x-\mu}{\sigma} - exp\left[-\frac{x-\mu}{\sigma}\right]\right], x \in \mathbb{R}$$

The Frechet distribution (*u element R, o > 0, alpha > 0*):

$$f(x) = \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-\alpha-1} exp\left[-\left(\frac{x-\mu}{\alpha}\right)^{-\alpha}\right], x > 0$$

The reversed Weibull distribution (*u element R, o > 0, alpha > 0*):

$$f(x) = \frac{\alpha}{\sigma}\left(\frac{\mu-x}{\sigma}\right)^{\alpha-1} exp\left[-\left(\frac{\mu-x}{\sigma}\right)^{\sigma}\right], x < \mu$$

Similarly, if $M = min\{X_1, ..., X_n\}$, we have that

$$\begin{aligned}F_M(m) &= \mathbb{P}(M \leq m) = 1 - \mathbb{P}(M > m)\\ &= 1 - \mathbb{P}(X_1 > m, \dots, X_n > m)\\ &= 1 - \mathbb{P}(X_1 > m)^n = 1 - (1 - F_X(m))^n\end{aligned}$$

Remark: It turns out that, when M is suitably shifted and scaled, there are again essentially three possibilities as $n \to$ *infinite*, being the distribution of $Y = -X$, where X is one of the three listed for the largest extreme value.

Summary

- Law of Large Numbers: statement, weak, strong
- Central Limit Theorem: statement, approximation of sums via CLT, examples
- Extreme Value Distributions: calculation (finite n), limiting behaviour (statement)

Statistics, Likelihood, and Estimation

Sums and Extremes of Independent Random Variables

- Law of Large Numbers; statement, weak, strong
- Central Limit Theorem; statement, approximation of sums via CLT, examples
- Extreme Value Distributions; calculation (finite n), limiting behaviour (statement)

Statistics

Data x is viewed as the outcome of a random variable X described by a probabilistic model. Usually, model is specified up to a (multidimensional) parameter: $X \sim F(.;\theta)$ for some element in Θ. In classical (frequentist) statistics, purely concerned with the model and in particular with the parameter $\emptyset$.
For example, given data, we may wish to

- estimate the parameter,
- perform statistical tests on that parameter, or
- validate the model

In Bayesian statistics, concerned with distribution of parameter $\theta \sim F(\theta)$.

Any real- or vector-valued function of data x or X is called a **statistic** of the data.
For example, the sample mean is a statistic:

$$T = T(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

given an outcome of X, or as a random variable

$$T = T(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Often, we will view data as a series of independent outcomes from the same random experiment: $X = (X_1, ..., X_N)$, where $X_1, ..., X_N$ are iid from $F(.;\theta)$. $\{X_1, ..., X_N\}$ is called a **random sample** (from $F(.;\theta)$ or from X).
Therefore, the joint cdf of a random sample is given by

$$F(x;\theta) = \prod_{k=1}^{N} F(x_k;\theta)$$

and so the joint pdf/pmf is of the same form:

$$f(x;\theta) = \prod_{k=1}^{N} f(x_k;\theta)$$

Likelihood

When viewed as a function of θ, then point pdf/pmf of a random sample is called the **Likelihood**:

$$L(\theta;x) = f(x;\theta)$$

The (natural) logarithm of the likelihood

$$l(\theta;s) = lnL(\theta;x)$$

is called the **log-likelihood**

Likelihood Example

Example: Model $X_{1}, {}_{2}, ..., X_N \sim iid\ Bin(m, p)$; m known, p unknown, in $\Theta = (0, 1)$
pmf:

$$f(x;p) = \binom{m}{x} p^x (1-p)^{m-x}, x \in \{0, 1, \ldots, m\}$$

Therefore, the likelihood can be written as

$$L(p;X) = \prod_{i=1}^{N} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}$$

$$= p^{\sum_{i=1}^{N} x_i}(1-p)^{Nm-\sum_{i=1}^{N} x_i} \prod_{i=1}^{N} \binom{m}{x_i}$$

Maximum Likelihood Estimation

How do we find "good" estimators for model parameters? Given data and a parametric model, how to find a member of that family (point estimate) from which the data is "most likely" to have come?
Given data x, one approach is to maximize the likelihood in θ – that is, find

$$\hat{\theta} \in \Theta$$

for which

$$L(\hat{\theta};x) \geq L(\theta;x), \theta \in \Theta$$

A maximizer

$$\hat{\theta} \equiv \hat{\theta}(x)$$

of L is called a **maximum likelihood estimate** (MLE). The corresponding random variable &hat;θ(X) is called a maximum likelihood estimator (also MLE).

Remark: A maximiser of *l* equivalent to a maximiser of *L*

ML Estimation Example: Binomial Probability

Example: Continuing our example, recall that we found

$$L(p; X) = \prod_{i=1}^{N} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}$$

$$= p^{\sum_{i=1}^{N} x_i}(1-p)^{Nm-\sum_{i=1}^{N} x_i} \prod_{i=1}^{N} \binom{m}{x_i}$$

How do we find an MLE?

Maximisation Strategy: Since *L* is a continuous function of *p*, find *p* such that

$$\frac{d}{dp}L(p; x) = 0$$

Working directly with *L* appears cumbersome; obtain the log-likelihood and work with that instead.

Taking the natural logarithm of *L*, we obtain the log-likelihood:

$$l(p; X) = ln(p) \sum_{i=1}^{N} x_i ln(1-p) \left(Nm - \sum_{i=1}^{N} x_i\right) + \sum_{i=1}^{N} ln \left(\binom{m}{x_i}\right)$$

First Derivative with respect to *p*:

$$\frac{d}{dp}l(p; X) = \frac{1}{p}\sum_{i=1}^{N} x_i - \frac{1}{1-p}\left(Nm - \sum_{i=1}^{N} x_i\right)$$

Set to zero and rearrange to find critical point:

$$(1-p)\sum_{i=1}^{N} x_i = p\left(Nm - \sum_{i=1}^{N} x_i\right)$$

Unique solution:

$$\hat{p} = \frac{1}{Nm}\sum_{i=1}^{N} x_i$$

What type of critical point is this?

Second Derivative with respect to *p*:

$$h(p) = \frac{d^2}{dp^2}l(p; X) = -\frac{1}{p^2}\sum_{i=1}^{N} x_i - \frac{1}{(1-p)^2}\left(Nm - \sum_{i=1}^{N} x_i\right) < 0$$

Therefore &hat;p is a local maximiser.

Moreover, *l(p;X) → &infty;* as *p → 0* or *p → 1* (boundary of Θ). Thus &hat;p is in fact a global maximiser.

Therefore, we have the Maximum Likelihood Estimator:

$$\hat{p} = \frac{1}{Nm}\sum_{i=1}^{N} X_i$$

Summary

- Statistics; definition, example
- Likelihood and log-likelihood; definition, binomial example
- Maximum Likelihood Estimation; definition, examples, bias, consistency