Prob and Stat with R

Roudranil Das

PBSR Aug-Nov 2022
Roll No: MDS202227
roudranil@cmi.ac.in

September 15, 2022

Contents

1 August 08, 2022

1.1 Random sampling and randomised controlled clinical trials

If clinical trials are not randomised, then what problems are there?

Problems are mainly ethical or there might be a judgemental bias in choosing which participant is in which group.

Ideally we would want all data/population groups to be proportionately or equally represented in all groups. This is impossible ot do by ourselves. In this, Fisher's suggestion was to assign them by tossing an unbiased coin. By the time we are done, the proportions are taken care of. We can also do the trials across multiple phases.

1.2 Sampling from social media data

Social media is **voluntary response survey methodology** (which always gives you an extremely biased view of the situation.) For good quality of data we should survey it properly, and although randomised survey methodology is generally expensive, the data it provides is usually much more representative of the population.

2 August 10, 2022

2.1 Social media data cant be used

In last class, we had seen how social media data cannot be used in order to predict political opinion (check **"The Great Hack"**). This is because this data is generally **oversampled**, that is some subset of the population is overrepresented in this population.

2.2 Hands on with R

Check the R markdown file. Some commands and stuff about R:

- While indexing, putting a negative integer removes that entry.
- `rbind.data.frame` binds rows of a data frame to an existing data frame.
-

3 August 17, 2022

3.1 Randomization

To understand the concept of randomization, note the issue in clinical trials again. We need randomised assignment of participants in the two groups. Check the R markdown file. Say we have created two groups of control and experiment, and both samples have very similar charecteristics across the board. Now if the experiment group shows vastly different results compared to the control group, we will know that whatever it is that we were testing has a really good effect as there was no inherent bias in how the groups were formed.

4 August 22, 2022

4.1 Random Variable

What is a random variable?

Usual definitions: function from the sample space to $\mathbb{R}$. Example: an accident is an random event. Following that we can have the following random variables:

(i) dead or alive (binary, discrete)

(ii) bill amount (continuous)

(iii) number of days in rehab (discrete, non binary)

$$X : \Omega \longrightarrow \mathbb{R} \tag{1}$$

We are associated in probability associated with random variable. Convention: $A, B, C, D \rightarrow$ *Random events*, $X, Y, Z, W \rightarrow$ *Random variables*. If they are in small letters, they are either observed events or realised random variables.

Say we want to try and introduce the idea of convergence in this probability. So we would need $0 < p(x) < 1$. Moreover since the sum of all probabilities should be 1, the infinite series with sum of probabilities should be convergent. Say we are dealing with discrete random variables. Then

$$\mathrm{P}[X \leq x] = \sum_{n=0}^{x} p(n) \rightarrow c < \infty \text{ as } x \rightarrow \infty$$

.

Here this following sequence will be very handy:

$$S(n) = \frac{\lambda^n}{n!}$$

Note that

$$\sum_{n=0}^{\infty} S(n) = e^{\lambda}$$

We define, $p(x) = \frac{S(x)}{e^{\lambda}} = e^{-\lambda}\frac{\lambda^x}{x!}$. Thus we have a probability distribution and the corresponding pmf (of ***Poisson*** distribution).

$$\boxed{\mathrm{P}[X \leq x] = \sum_{n=0}^{x} e^{-\lambda}\frac{\lambda^n}{n!}} \tag{2}$$

Cumulative distribution function ↑

4.1.1 Continuous random variable

For probability distribution for continuous r.v.,

$$\mathrm{P}[X \leq x] = \int_{-\infty}^{x} f(x)dx$$

5 August 24, 2022

5.1 Continuous Random Variable (contd.)

From Fisher's Definition, if we have data in bins and relative frequency, then area of bins is relative frequency which is an approximation of the probability. Suppose we have a histogram, y-axis being density and x-axis being the bins. thn $\mathrm{P}[c < X < d] = B_1 + B_2 + B_3 + B_4 = \sum_{i=1}^{4} B_i = \sum_{i=1}^{4} d_i \times \Delta_i$. Thus we will sort of need to approximate the density d_i by some function $f(x_i)$. And we wll need $f(x_i) > 0$.

Thus,

$$\begin{aligned}\mathrm{P}[a < X < b] &= \sum_{i=1}^{k} f(x_i)(x_i - x_{i-1}) \\ &= \int_a^b f(x)dx \text{ (as } \Delta \rightarrow 0)\end{aligned}$$

This is the probability density function, with the usual properties. Thus by mathematical construction, $\mathrm{P}[X = a] = 0$.

$F_X(a) = \mathrm{P}[X < a]$ (cdf).

5.2 Moments

Say from a distribution, we have $x_1, x_2, \ldots, x_n$ (SRSWR). Then define, $\bar{x} = \frac{\sum x_i}{n}$ (sample average). We can sort of rewrite this as $\bar{x} = \sum x_i w_i$ where $w_i = \frac{1}{n} \forall i$ Then

$$\bar{x} = \sum x_i p(x_i) = \mathbb{E}[X]$$

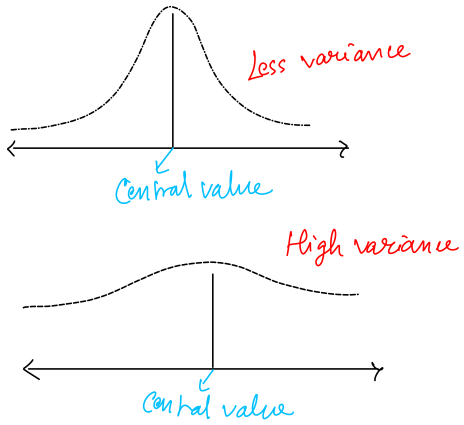
For continuous,

$$\mathbb{E}[X] = \int x f(x) dx$$

This is First order raw moments. N^{th} order raw moments will be $\mathbb{E}[X^n]$.

6 September 1, 2022

6.1 Central moments



$$\mathbb{V}(X) = \int (x - \mu)^2 f(x) dx = \mathbb{E}[x - \mu]^2 \tag{3}$$

This is sort of a probability version. For sample, if the random sample drawn is $\{x_1, x_2, \ldots, x_n\}$ then the sample mean is $\bar{x} = \frac{1}{n} \cdot \sum x_i$. Then the sample variance is $s^2 = \frac{1}{n} \cdot \sum (x_i - \bar{x})^2$. Or if we take the unbiased estimator for $s^2 = \frac{1}{n-1} \cdot \sum (x_i - \bar{x})^2$.

6.2 Moment generating function

$\mathbb{E}[e^{tX}] = \int e^{tX} f(x) dx$ is the moment generating function. Specifically, $\frac{\partial \mathbb{E}[e^{tX}]}{\partial t} = \mathbb{E}[X]$ at $t = 0$.

Why? Because by expanding $e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \cdots$ and taking expectation we get $\mathbb{E}[e^{tX}] = \mathbb{E}[1 + tX + \frac{(tX)^2}{2!} + \cdots]$. Hence after after differentiating wrt t and setting $t = 0$, we get the expectation. Higher order partital derivatives give the other moments.

6.3 Conditional Probability

Say we have two events A and S. Then

- $\mathbb{P}(A) \to$ marginal probability of A
- $\mathbb{P}(A \mid S) \to$ conditional probability of A given S

If $\mathbb{P}(A) = \mathbb{P}(A \mid S)$ then the two events are independent.

For random variables, $\mathbb{P}(X) \to$ prob dist of X. $\mathbb{P}(X \mid Y = y) \to$ cond prob dist of X given $Y = y$. $\mathbb{P}(X, Y) \to$ Joint prob dist. If $\mathbb{P}(X, Y) = \mathbb{P}(X) \cdot \mathbb{P}(Y)$, then the two random variables are independent.

7 September 5, 2022

7.1 SRSWR and SRSWOR

Say we have an urn, with multiple balls of different colours. Then say we draw a sample of size 3, (SRSWR), and record the colours. Also we do this SRSWOR (another sample of size = 3).

- **In the first draw if we see blue, what is the probability we will see a blue ball in the next draw?**

$$\mathbb{P}(B_2 \mid B_1) = \begin{cases} \frac{1}{5}, \text{SRSWR} \\ 0, \text{SRSWOR} \end{cases}$$

In SRSWR, draws are independent. In SRSWOR draws are not independent.

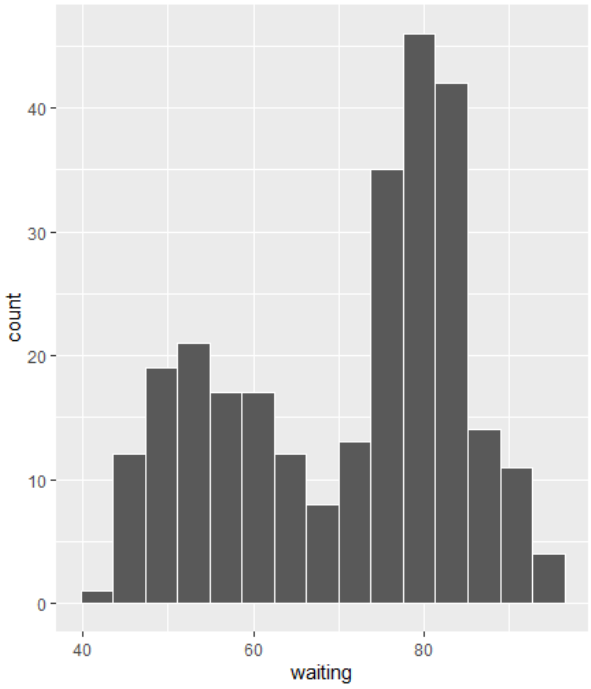
There are two kinds of data set we will need to work with in order to understand the two simple random sampling schemes.

1. Cross-sectional Data set: Essentially SRSWR. If we have an experiment going on for a long time, then at one time point, the data across all the attributes is cross-sectional data set.
2. Time-series Data set (or transversal data set): Series of data indexed by time points. Weather data, proft and loss data across a long time. And because information from today affects information of the future, the observations are not independent. Hence, the data is SRSWOR.
3. Longitudinal Data set (panel data): Track the same sample at different points of time. For any one sample, the data will be like a time series. But at any time point, the samples will behave like a cross-sectional data set.

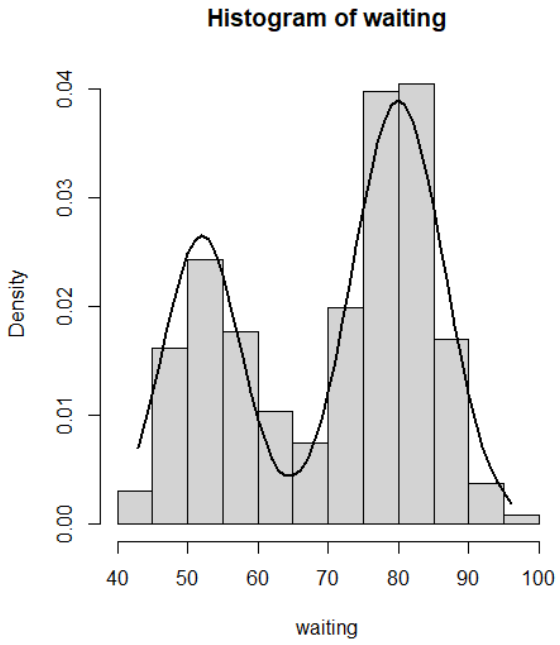
8 8 September, 2022

8.1 Mixture models

We take multiple probability distributions, then their convex combination is another probability distribution. This can be used to approximate complex distributions with different properties. For Example:



This was approximated by:



This is $p \cdot N(52, 5.5^2) + (1-p) \cdot N(80, 6.5^2)$, where $0 < p < 1$.

Say we want to compute the MLE of parameters of a bimodal mixed distribution, like this one. Then we define the negative log likelihood and use the `optim` function to find the optimum parameters.

```
NegLokLikeMix <- function(theta, data){
    mu1 = theta[1]
    sigma1 = exp(theta[2]) # because optim runs it on the real line, but sigma needs to be
        positive

    mu2 = theta[3]
    sigma2 = exp(theta[4])

    p = exp(theta[5])/(1+exp(theta[5])) # logit transformation

    n = length(data)

    l = 0
    for(i in 1:n){
        l = l + log(p*dnorm(data[i], mean = mu1, sd = sigma1) + (1-p)*dnorm(data[i], mean =
            mu2, sd = sigma2))
    }
    return(-l)

}
theta_initial = c(52,8,80,8,0.5)
optim(theta_initial,
        NegLokLikeMix,
        data=waiting,
        control = list(maxit = 1500))
```

What we are doing here is simple likelihood optimisation.

Say now we have the optimal parameters. So we say the model has been fit. Now what?

9 12 September, 2022

9.1 Datasets analysed

1. Football Dataset :-
   - Poisson model
   (a) method of moments
   (b) Maximum likelihood estimation to find the unknown paramenter of the model λ
   - Using Poisson Model we did some inference.
2. Old faithful geyser dataset :-
   - $\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}$ data points - We had a mixture model.
   (a) used the likelihood method to estimate the 5 parameters (check the rmd)

Estimate $\hat{\theta}$ is not going to be exactly θ_0 (true parameter). We dont actually know this θ_0. We have a dataset $D = \{X_1, X_2, \ldots, X_n\}$ and it contains some information about θ_0.

9.2 Sampling distributions

Suppose we are doing a survey in order to find out the avg income of an area. So we take random samples and calculate their sample means (given that we have ∞ resources).

1. $[x_1, x_2, \ldots, x_n] = \tilde{x}_1, \hat{x}_1$
2. $[x_1, x_2, \ldots, x_n] = \tilde{x}_2, \hat{x}_2$
   .
   .
   .

1000. $[x_1, x_2, \ldots, x_n] = \tilde{x}_{1000}, \hat{x}_{1000}$

Say we make an histogram with $\hat{x_r}$. Then we will get an idea of their probability distribution. This is the sampling distribution of the sample means.

Sampling distribution of an estimator (not neccessarily sample mean) tells us that what is the expected error that estimator can make, and what is the unlikely error for that estimator.

Say for a bell shaped distribution, mean = median. Then

Q: Shall we use sample mean or sample median?

We build the sampling distribution of both, and take the one with less standard error. Or its better if we make an educated choice.

When we increase the sample size the sampling distribution becomes sharper towards the true value (at the limiting case it will become a degenerate distribution). And if we take a small distribution, then it will become flatter.

9.3 Distribution under transformation

(read CSBGR chap2 pgs 47 - 68)

Exercise 9.3.1. $X \sim N(\mu, \sigma^2)$. Calculate the moment generating function of X.

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} f(x)dx = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Now say $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $S = \sum X_i$, $\bar{X} = \frac{S}{n}$. Then what is the distribution of S? To find that we find $M_S(t)$.

$$M_S(t) = \mathbb{E}[e^{tS}] = \mathbb{E}\left[e^{t\sum X_i}\right]$$

$$= \mathbb{E}\left[\prod e^{tX_i}\right] = \prod \mathbb{E}\left[e^{tX_i}\right]$$

We could write the last line because of the iid. From there we can say that $S \sim N(n\mu, n\sigma^2)$.

Exercise 9.3.2. Show $\bar{X} \overset{iid}{\sim} N(\mu, \frac{\sigma^2}{n})$

Proof. Method 1: Using the independence.

Given that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$,
$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\mathbb{E}\left[\sum X_i\right] = \mu$
$\mathbb{V}\left(\bar{X}_n\right) = \frac{1}{n^2}\sum \mathbb{V}\left(X_i\right) = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}$.
$\therefore \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Better method would be with mgf. □

10 14 September, 2022

10.1 Central Limit Theorem

Lindeberg-Levy's Central Limit Theorem (CLT)

Suppose $\{X_1, X_2, \ldots, X_n\}$ are iid random variables such that:

- $\mathbb{E}[X_i] = \mu$
- $\mathbb{V}(X_i) = \sigma^2 < \infty$ (automatically this means that the mean exists)

Then as $n \to \infty$,
$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{\mathcal{L}} N(0,1)$

What this means is that, if we have large enough sample size, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$. Most importantly, this theorem does not assume any sort of underlying distribution. We only need their mean and variance to exist.

Sampling distribution of sample proportion will approximately follow normal if we can ensure large enough sample sizes.

$$\hat{p} = \bar{X} \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Previously we checked if the underlying distribution is Gamma, then the sampling distribution of the sample mean is also Gamma. But due to CLT, as sample size grows, the exact sampling distribution which is Gamma will start behaving like Gaussian.

Typically we call this $P_n \xrightarrow{\mathcal{L}} P$ or $P_n \xrightarrow{\mathcal{D}} P$: convergence in law or convergence in distribution.

CLT only talks about sampling distribution of sample mean.

10.2 Probability distributions under transformations

If X is a random variable with cdf $F_X(x)$, $Y = g(X)$ and g is deterministic. Then Y is also a random variable.

Example. X is rv. $Y = \log(X)$ is a rv.

$P(Y \in A) = P(g(X) \in A)$
Say, $X \in \mathcal{X}$: sample space (range of values of X). Then,

$$g(x) : \mathcal{X} \longrightarrow \mathcal{Y}$$

We assume that inverse mapping of g also exists and we denote it at g^{-1}. For $A \subseteq \mathcal{Y}$

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$$

The 3 axioms of Kolmogorov holds for $P(Y \in A)$ under the transformation $Y = g(X)$.