Module 4: Statistical inference (I)

Yaqi Shi

July 16, 2024

Outline

This module we will review

- Basics of probability
- Fundamental concepts in inference

Probability distributions

- In statistics, we try to draw conclusions about a larger population from a sample of observations.
- We use mathematical models to capture probabilistic behavior of a population.
- This behavior is modeled using probability distributions.

Density/Distribution functions

Definition (Cumulative Distribution Function)

$$F_X(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$$

Density/Distribution functions (cont'd)

Definition (Probability Mass Function)

For a discrete RV, the probability mass function (PMF) is:

$$f_X(x) = P(X = x) \quad \forall x \in \mathbb{R}$$

Definition (Probability Density Function)

For a continuous RV, the probability density function (PDF) is:

$$f_X(x) = \frac{\partial}{\partial t} F(t)\bigg|_{t=x}$$

So $F_X(x) = \int_{-\infty}^{x} f_X(t) dt \forall x \in \mathbb{R}$.

Note that $f_X \geq 0$ for $\forall x$, and thus F_X is an increasing function.

Expectation and Variance

Definition (Expectation)

A measure of central tendancy (a weighted average of the values of X)

$$E[X] = \sum_{x \in S} xP(X = x) \text{ for discrete RV taking values from } S$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \text{ for continuous RV}$$

Definition (Variance)

A measure of the spread of a distribution

$$\text{Var}(X) = \sum_{x \in S} (x - E[X])^2 P(X = x) \text{ for discrete RV}$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx \text{ for continuous RV}$$

Discreate random variable

A discrete random variable has a countable number of possible values.

Bernoulli and Binomial random variable

- Consider the event of flipping a (possibly unfair) coin.
- $Y \in \{0, 1\}$ represents success and failure.
- Suppose we only flip the coin once,
  - We can express $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$
- Bernoulli distribution

$$P(Y = y) = p^y (1 - p)^{1-y} \quad \text{for} \quad y = 0, 1$$

- If we flip the coin n times,
- Binomial distribution

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad \text{for} \quad y = 0, 1, \ldots, n$$

Binomial distributions with different values of n and p

If $Y \sim \text{Binomial}(n, p)$, then $\mathrm{E}(Y) = np$ and $\text{SD}(Y) = \sqrt{np(1-p)}$.

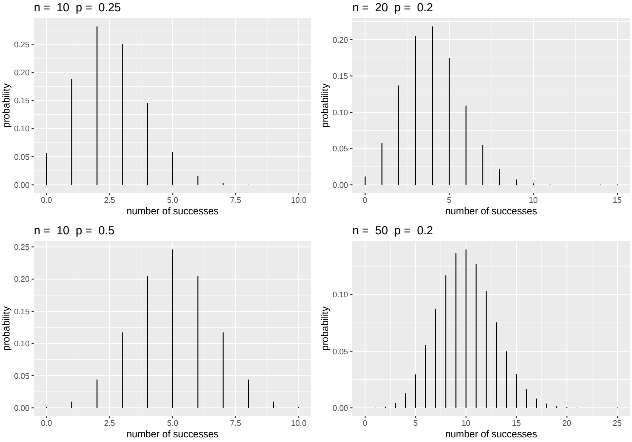


Figure 1: Binomial distributions with different values of n and p.

How to generate in R?

All common distributions have four functions in R:

- Density
  `dbinom(x, size, prob)`
- Distribution function
  `pbinom(q, size, prob)`
- Quantile function
  `qbinom(p, size, prob)`
- Random generaation
  `rbinom(n, size, prob)`

Not sure? Using `?` with any of the four functions, e.g. `?qbinom`

Example of binomial distribution computing

Question: While taking a multiple choice test, a student encountered 10 problems where she ended up completely guessing, randomly selecting one of the four options. What is the chance that she got exactly 2 of the 10 correct?

Example of binomial distribution computing

Question: While taking a multiple choice test, a student encountered 10 problems where she ended up completely guessing, randomly selecting one of the four options. What is the chance that she got exactly 2 of the 10 correct?

Answer: Knowing that the student randomly selected her answers, we assume she has a 25% chance of a correct response.

$$P(Y = 2) = \binom{10}{2} (.25)^2 (.75)^8 = 0.282$$

Example of binomial distribution computing

Question: While taking a multiple choice test, a student encountered 10 problems where she ended up completely guessing, randomly selecting one of the four options. What is the chance that she got exactly 2 of the 10 correct?

Answer: Knowing that the student randomly selected her answers, we assume she has a 25% chance of a correct response.

$$P(Y = 2) = \binom{10}{2} (.25)^2 (.75)^8 = 0.282$$

R computing:

```
dbinom(2, size = 10, prob = .25)
```

```
## [1] 0.2815676
```

Continuous random variable

A continuous random variable can take on an uncountably infinite number of values. Given a pdf $f(y)$,

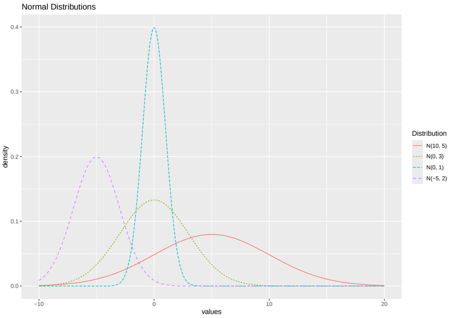
$$P(a \leq Y \leq b) = \int_a^b f(y)dy$$

Properties:

- $\int_{-\infty}^{\infty} f(y)dy = 1$.
- For any value y, $P(Y = y) = \int_y^y f(y)dy = 0$.
  $P(y < Y) = P(y \leq Y)$.

Example of Continuous Distribution (Normal)

- The normal distribution is a very important distribution because:
  - A lot of things look normal
  - Analytically tractable
  - Central limit theorem
- $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- Characterized by mean, μ, and variance, σ^2.



How to Generate Samples from Normal Distribution

The following commands are for a normal random variable with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$,

- To calculate the probability density function at a value x, `dnorm(x,mu,sigma)`
- To calculate the cumulative distribution function at a value x, `pnorm(x,mu,sigma)`
- To generate a size m sample from the normal distribution, `rnorm(m,mu,sigma)`
- Note that the third argument is the **square root of the variance**, this is because the R function for normal distribution asks for the standard deviation, which is defined as the square root of the variance

Some probability distributions in R

Continuous

- Normal (`?rnorm`)
- Uniform (`?runif`)
- Beta (`?rbeta`)
- Chi-sq (`?rchisq`)
- Exponential (`?rexp`)
- t (`rt`)
- F (`?rf`)
- Logistic (`?rlogis`)
- Lognormal (`?rlnorm`)

Discrete

- Poisson (`?rpois`)
- Binomial (`?rbinom`)
- Geometric (`?rgeom`)
- Negative Binomial (`?rnbinom`)
- Multinomial (`?rmultinom`)

Empirical vs. Theoretical CDF

In statistics, an empirical distribution function is the distribution function associated with the empirical measure of a sample.

- Theoretical CDF

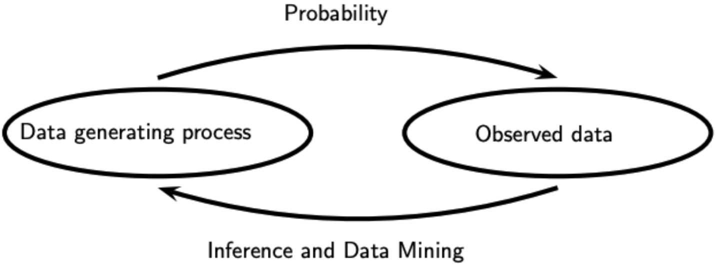
$$F_X(k) = \Pr(X \leq k)$$

- Empirical CDF

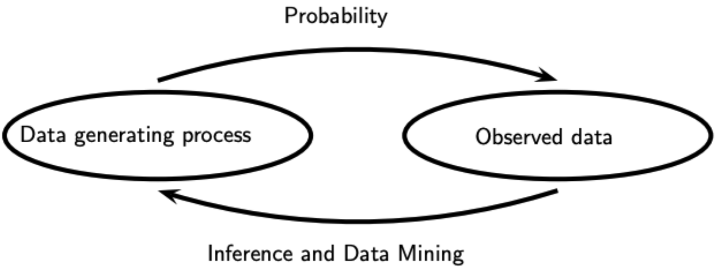
$$\hat{F}_n(k) = \frac{\text{number of elements in the sample } \leq k}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq k}$$

where $X_1, \ldots, X_n$ make up some random sample from the underlying distribution.

Probability and inference



Probability and inference



- Probability: Given a data generating process, what are the properties of the outcomes?
- Statistical inference: Given the outcomes, what can we say about the process that generated the data?

Parametric vs. Nonparametric models

- Statistical model $\mathfrak{F}$: a set of distributions (or densities or regression functions)

Parametric vs. Nonparametric models

- Statistical model $\mathfrak{F}$: a set of distributions (or densities or regression functions)

- Parametric model: a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$$

  where θ is an unknown parameter (or vector of parameters) that can take values in the parameter space Θ.

  - e.g. Normal distribution, a 2-parameter model with density as $f(x;\mu,\sigma)$

Parametric vs. Nonparametric models

- Statistical model $\mathfrak{F}$: a set of distributions (or densities or regression functions)

- Parametric model: a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters

$$\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$$

  where θ is an unknown parameter (or vector of parameters) that can take values in the parameter space Θ.

  - e.g. Normal distribution, a 2-parameter model with density as $f(x; \mu, \sigma)$

- Nonparametric model: a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters
  - e.g. $\mathfrak{F}_{\mathrm{ALL}} = \{\text{ all } \mathrm{CDF}'s\}$ is nonparametric.

Frequentist and Bayesian

- Frequentist: statistical methods with guaranteed frequency behavior
- Bayesian: statistical methods for using data to update beliefs

Point estimation

- Providing a single “best guess” of some quantity of interest
- Notations
  - Parameter θ: fixed, unknown quantity
  - Point estimator $\hat{\theta}$: depends on data, random variable

Point estimation

- Providing a single "best guess" of some quantity of interest
- Notations
  - Parameter θ: fixed, unknown quantity
  - Point estimator $\hat{\theta}$: depends on data, random variable

Definition (Point estimator)

Let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g\left(X_1, \ldots, X_n\right)$$

- What is a good point estimate?

MSE

- Definition:

$$\mathrm{MSE} = \mathbb{E}_\theta \left(\widehat{\theta}_n - \theta\right)^2$$

- No uniformly best estimator in terms of MSE
- It is NOT possible to have an estimator that is uniformly the best.

Bias and Variance

- Bias

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta$$

- Variance

$$\mathrm{Var}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n - \mathbb{E}\theta\right)^2$$

- Theorem

$$MSE = bias^2 + Var$$

Unbiasedness

- Definition

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta = 0$$

- Unbiasedness is a small sample (finite sample) property
- An unbiased estimator may not exist
- An unbiased estimator is not necessarily a good estimator

Consistency

- Definition

$$\widehat{\theta}_n \xrightarrow{\text{P}} \theta$$

- It is possible to be unbiased but not consistent.
- It is possible to be consistent but not unbiased.

Asypototic unbiasedness

- Definition

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta \to 0, \text{ as } n \to \infty$$

- It is possible to be asypototically unbiased but not consistent.
- It is possible to be consistent but NOT asymptotically unbiased.
- Sufficient conditions: $MSE \to 0$.

Resources

This tutorial is based on

- Havard Biostatistics Summer Pre Course [link]
- "Beyond Multiple Linear Regression" by Paul Roback and Julie Legler [link]