Module 4: Statistical inference (I)

Siyue Yang

05/10/2022

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:

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

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 $\mathrm{SD}(Y) = \sqrt{np(1-p)}$.

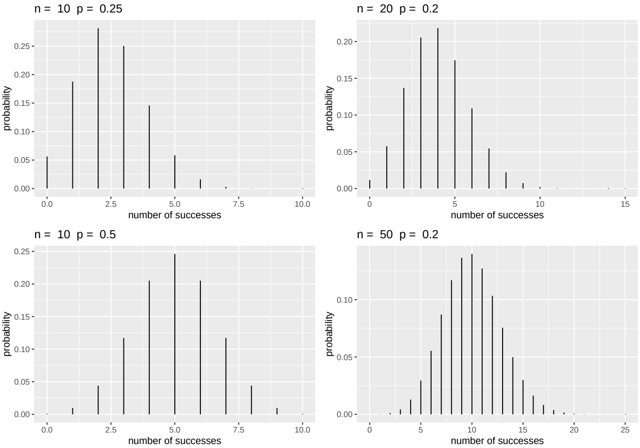


Figure 1: Binomial distributions with different values of n and p.

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.

$$P(Y = 2) = \left(\begin{array}{c} 10 \\ 2 \end{array} \right) (.25)^2 (.75)^8 = 0.282$$

R computing:

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

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

Geometric random variables

- Suppose we are to perform independent, identical Bernoulli trials until the first success.
- If we wish to model Y, the number of failures before the first success
- Geometric distribution

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

Geometric distributions with $p = 0.3,\ 0.5$ and 0.7

If $Y \sim \text{Geometric}(p)$, then $E(Y) = \frac{1-p}{p}$ and $SD(Y) = \sqrt{\frac{1-p}{p^2}}$.

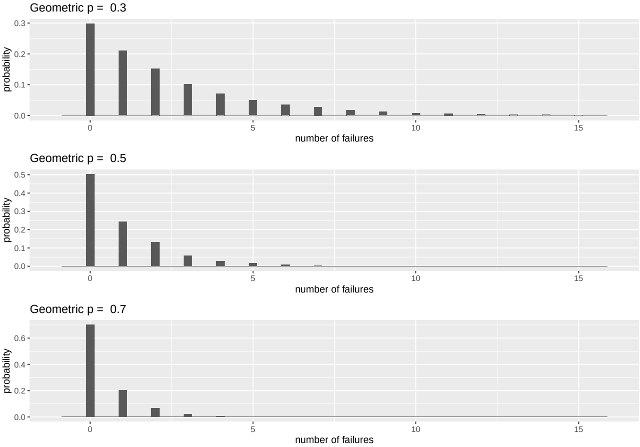


Figure 2: Geometric distributions with $p = 0.3, 0.5$ and 0.7.

Negative binomial random variable

- If we were to carry out multiple independent and identical Bernoulli trails until the r^{th} success occurs.
- Y, the number of failures before the r^{th} success
- Negative binomial distributions

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

- When $r = 1$, the geometric distribution is a special case of negative binomial distribution.

Negative binomial distributions with different p and r

If $Y \sim \mathrm{NB}(r, p)$ then $E(Y) = \frac{r(1-p)}{p}$ and $SD(Y) = \sqrt{\frac{r(1-p)}{p^2}}$.

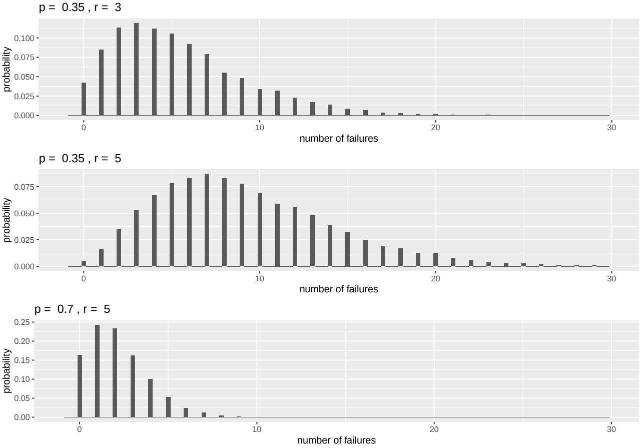


Figure 3: Negative binomial distributions with different values of p and r.

Hypergeometric random variable

- Bernoulli process assumes the probability of a success remained constant across all trials.
- What if this probability is dynamic?

Hypergeometric random variable

- Bernoulli process assumes the probability of a success remained constant across all trials.
- What if this probability is dynamic?

- Suppose we wanted to select n items **without replacement** from a collection of N objects, m of which are considered successes?
- The probability of selecting a "success" depends on the previous selections.
- Y, the number of successes after n selections
- Hypergeometric random variable

$$P(Y = y) = \frac{\binom{m}{y}\binom{N-m}{n-y}}{\binom{N}{n}} \quad \text{for} \quad y = 0, 1, \ldots, \min(m, n).$$

Hypergeometric distributions with m, N, and n

Y follows a hypergeometric distribution and we define $p = m/N$, then $E(Y) = np$ and $SD(Y) = \sqrt{np(1-p)\frac{N-n}{N-1}}$.

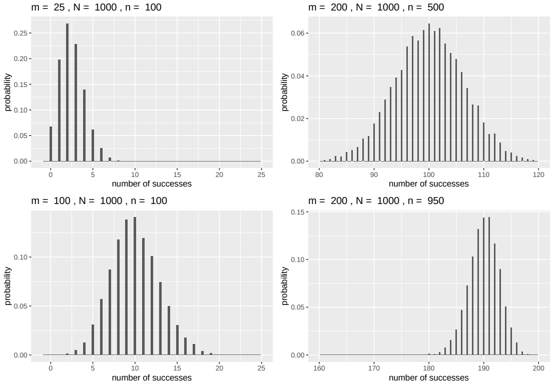


Figure 4: Hypergeometric distributions with different values of m, N, and n

Poisson random variable

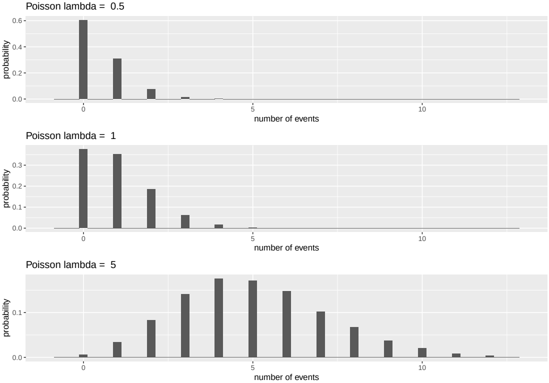
- In a Poisson process, we are counting the number of events per unit of time or space and the number of events depends only on the length or size of the interval.
- Y, the number of events
- Poisson distribution

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad \text{for} \quad y = 0, 1, \dots, \infty,$$

where λ is the mean or expected count in the unit of time or space of interest.

Poisson distributions with $\lambda = 0.5, 1,$ and 5

$E(Y) = \lambda$ and $SD(Y) = \sqrt{\lambda}$



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, $P(Y = y) = \int_y^y f(y)dy = 0$.
  $P(y < Y) = P(y \leq Y)$.

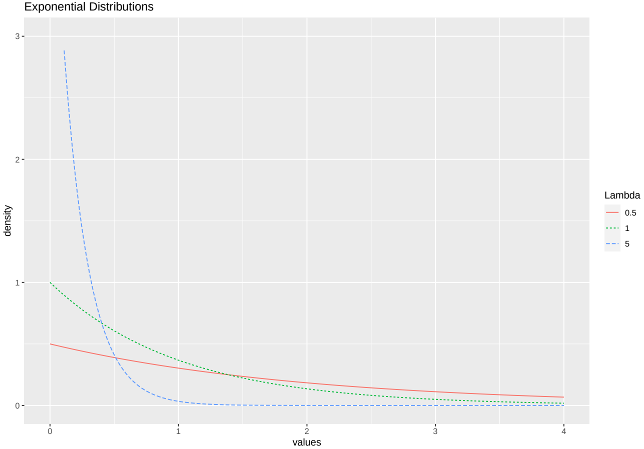
Exponential random variable

- Suppose we have a Poisson process with rate λ
- To model the wait time Y until the first event
- Exponential distribution

$$f(y) = \lambda e^{-\lambda y} \quad \text{for} \quad y > 0,$$

Exponential distributions with $\lambda = 0.5, 1,$ and 5

$E(Y) = 1/\lambda$ and $SD(Y) = 1/\lambda$



Gamma random variable

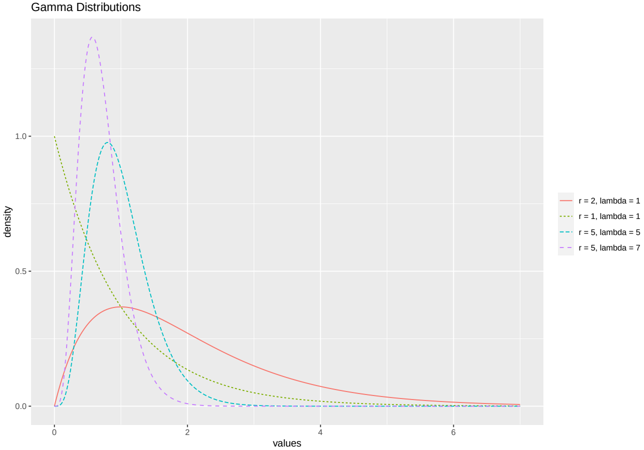
- Consider a Poisson process.
- Y, waiting time before 1 event occurrd, follows an exponential distribution.
- Y, waiting time before r events occurred, follows a gamma distribution.

$$f(y) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \quad \text{for} \quad y > 0$$

- When $r = 1$, the exponential distribution is a special case of gamma distribution.

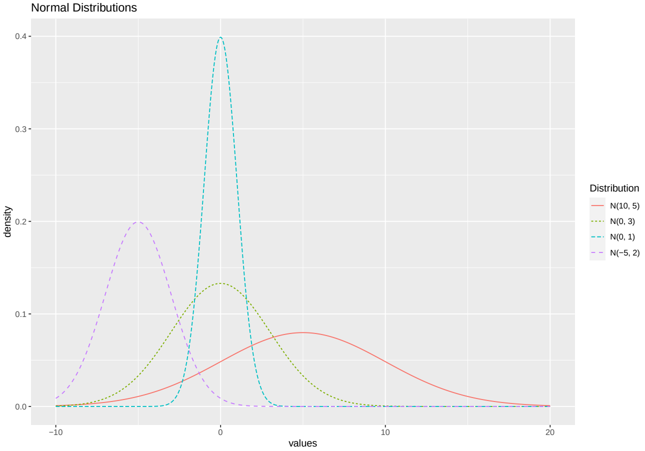
Gamma distributions with different values of r and λ

If $Y \sim \mathrm{Gamma}(r, \lambda)$ then $E(Y) = r/\lambda$ and $SD(Y) = \sqrt{r/\lambda^2}$.



Normal random variable

$Y \in N(\mu, \sigma^2)$, $\mathrm{E}(Y) = \mu$ and $\mathrm{SD}(Y) = \sigma$.



Beta random variable

We often use beta random variables to model distributions of probabilities bounded below by 0 and above by 1.

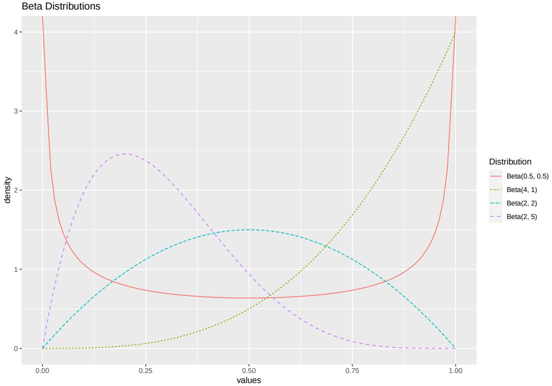
$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad \text{for} \quad 0 < y < 1$$

- If $\alpha = \beta = 1$, it follows a uniform distribution,

$$\begin{aligned} f(y) &= \frac{\Gamma(1)}{\Gamma(1)\Gamma(1)} y^{0}(1-y)^{0} \\ &= 1 \quad \text{for} \quad 0 < y < 1. \end{aligned}$$

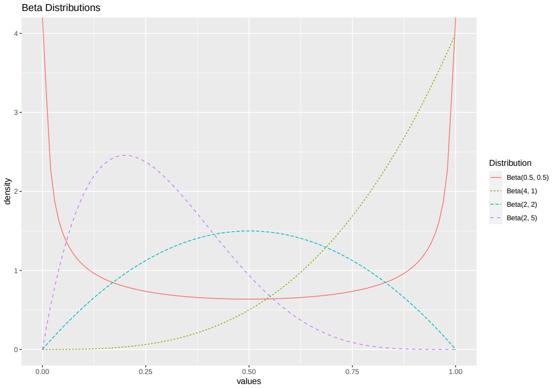
Beta distributions with different values of α and β

$Y \sim \mathrm{Beta}(\alpha, \beta)$, then $E(Y) = \alpha/(\alpha+\beta)$ and $SD(Y) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$.



Beta distributions with different values of α and β

$Y \sim \mathrm{Beta}(\alpha, \beta)$, then $E(Y) = \alpha/(\alpha + \beta)$ and $SD(Y) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$.



Note that when $\alpha = \beta$, distributions are symmetric. The distribution is left-skewed when $\alpha > \beta$ and right-skewed when $\beta > \alpha$.

Distributions used in testing

- χ^2 distribution
- t distribution
- F distribution

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

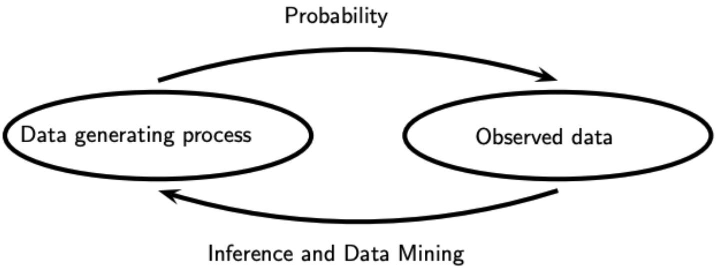
$$F_X(k) = \Pr(X \leq k)$$

- Empirical CDF

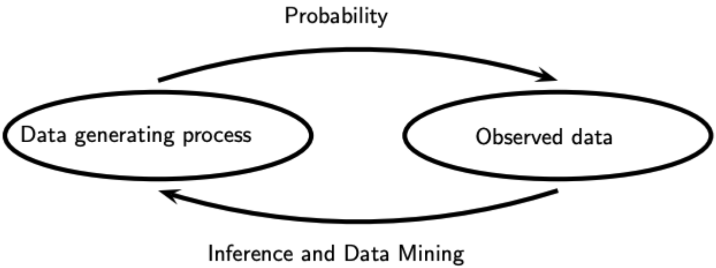
$$\hat{F}_n(k) = \frac{\text{number of elements in the sample } \leq k}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq k}$$

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

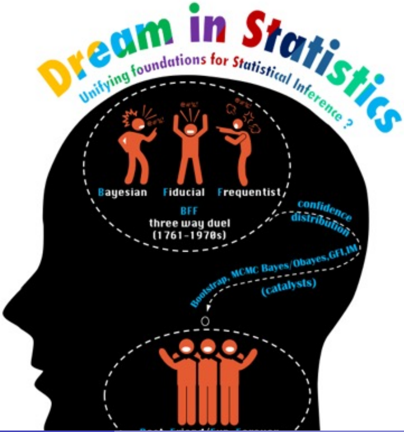
- Statistical model $\mathfrak{F}$: a set of distributions (or densities or regression functions)
- Parametric model: a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters

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

  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.

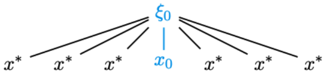
Frequentist, Bayesian, Fiducial inference (BFF)

- Frequentist: statistical methods with guaranteed frequency behavior
- Bayesian: statistical methods for using data to update beliefs
- Fiducial: statistical methods based on inverse probability without calling on prior probability distributions

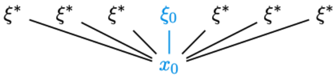


Difference: math details, interpretation, replication

- Frequentist: modeling collection of distributions $\mathcal{P} = \{P_\xi\}_{\xi \in \Xi}$
  - parameter ξ_0 fixed, data x replicated



- Bayesian: modeling one joint distribution $f(x \mid \xi) \cdot \pi(\xi)$
  - data x_0 fixed, parameter ξ replicated



- Fiducial: modeling data generating algorithm $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
  - data x & parameter ξ linked through DGA, auxiliary variable u replicated

Fundamental concepts in inference

- Point estimation
- Hypothesis testing
- Confidence sets

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 $X_1, \ldots, X_n$:

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

- Properties
  - Unbiasedness
  - Consistency
  - Efficiency

Point estimation (cont'd)

- Bias

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

- Consistency

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

- Standard error

$$\mathrm{se} = \text{se}\left(\hat{\theta}_n\right) = \sqrt{\mathbb{V}\left(\hat{\theta}_n\right)}$$

- Mean square error

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

Confidence sets

Definition (Confidence set)

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$\mathbb{P}_\theta (\theta \in C_n) \geq 1-\alpha \quad \forall \theta \in \Theta$$

- If θ is a vector, we use **Confidence sets** instead of **Confidence intervals**.
- In Frequentist, θ is fixed while C_n is random.
  - Confidence interval is not a probability statement about θ.
- In Bayesian, θ is random.
  - Bayesian interval refers to degree-of-belief probabilities.

Hypothesis testing

Definition (Hypothesis testing)

Suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis.

Hypothesis testing (cont'd)

Let X be a random variable, $\mathcal{X}$ be the range of X. We test a hypothesis by finding the rejection region $R \subset \mathcal{X}$,

$$\begin{aligned} X \in R \quad &\Longrightarrow \quad \text{reject } H_0 \\ X \notin R \quad &\Longrightarrow \quad \text{retain (do not reject) } H_0 \end{aligned}$$

Common form of R,

$$R = \{x : T(x) > c\}$$

where T is a test statistic and c is a critical value.

Hypothesis testing (cont'd)

- Type I error: Rejecting H_0 when H_0 is true
- Type II error: Retaining H_0 when H_1 is true

Definition (Power function)

The power function of a test with rejection region R is defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

The size of a test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

A test is said to have level α if its size is less than or equal to α.

Resources

This tutorial is based on

- Havard Biostatistics Summer Pre Course [link]
- "Beyond Multiple Linear Regression" by Paul Roback and Julie Legler [link]
- "Short course on Generalized Fiducial Inference" by Jan Hannig [link]

More resources: - BFF, Bayesian, Fiducial & Frequentist: http://bff-stat.org/about/