Math 408 - Mathematical Statistics

Lecture 1. ABC of Probability

January 16, 2013

Agenda

- Sample Spaces
- Realizations, Events
- Axioms of Probability
- Probability on Finite Sample Spaces
  - Example: B-day Problem
- Independent Events
- Summary

Sample Spaces, Realizations, Events

Probability Theory is the mathematical language for uncertainty quantification.

The starting point in developing the probability theory is to specify sample space = the set of possible outcomes.

Definition

- The **sample space** Ω is the set of possible outcomes of an "experiment"
- Points $\omega \in \Omega$ are called **realizations**
- **Events** are subsets of Ω

Next, to every event $A \subset \Omega$, we want to assign a real number $\mathbb{P}(A)$, called the probability of A. We call function $\mathbb{P} : \{\text{subsets of } \Omega\} \to \mathbb{R}$ a probability distribution.

We don't want $\mathbb{P}$ to be arbitrary, we want it to satisfy some natural properties (called axioms of probability):

1. $0 \leq \mathbb{P}(A) \leq 1$ (Events range from never happening to always happening)
2. $\mathbb{P}(\Omega) = 1$ (Something must happen)
3. $\mathbb{P}(\emptyset) = 0$ (Nothing never happens)
4. $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$ (A must either happen or not-happen)
5. $\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

Probability on Finite Sample Spaces

Suppose that the sample space is finite $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$.

Example:

If we toss a die twice, then Ω has $n = 36$ elements:

$$\Omega\{\ (i,j)\ :\ i,j = 1,2,3,4,5,6\}$$

If each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$, where $|A|$ denotes the number of elements in A.

Test question: What is the probability that the sum of the dice is 11?

Answer: $2/36$, since the are two outcomes that correspond to this event: $(5,6)$ and $(6,5)$.

In general, if Ω is finite and if each outcome is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

To compute the probability $\mathbb{P}(A)$, we need to count the number of points in an event A. Methods for counting points are called combinatorial methods.

Example: Birthday Problem

Suppose that a room of people contains n people.
What is the probability that at least two of them have a common birthday?

Assume that

- Every day of the year is equally likely to be a birthday
- There are 365 days in the year (disregard leap years)

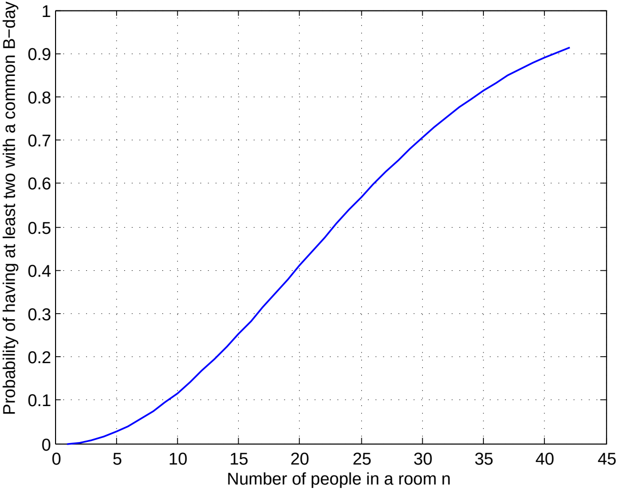
Then

- $\Omega = \{\omega = (x_1, \dots, x_n) \ : \ x_i = 1, 2, \dots, 365\}, \quad |\Omega| = 365^n$
- $A = \{\omega \in \Omega \ : \ x_i = x_j \text{ for some } i \neq j\}$
- $\bar{A} = \{\omega \in \Omega \ : \ x_i \neq x_j \text{ for all } i, j\}, \quad |\bar{A}| = 365 \times 364 \times \dots \times (365 - n + 1)$

$$\boxed{\mathbb{P}(A) = 1 - \frac{365 \times 364 \times \dots \times (365 - n + 1)}{365^n}}$$

n	$\mathbb{P}(A)$
4	0.016
23	0.507
32	0.753
42	0.91
56	0.988

Example: Birthday Problem



Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. We can formalize this useful notion of independence as follows:

Definition

Two events A and B are **independent** if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

Independence can arise in two distinct ways:

1. We explicitly assume that two events are independent. For example, in tossing a coin twice, we usually assume that the tosses are independent which reflects the fact that the coin has no memory of the first toss.
2. We derive independence of A and B by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. For example, in tossing a fair die, let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Are A and B independent?
   Yes! Since $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$, $AB = \{2, 4\}$,
   $\mathbb{P}(AB) = 1/3 = (1/2) \times (2/3)$

Examples

- Suppose that A and B are disjoint events, each with positive probability. Can they be independent?
  Answer: No! $\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0$, but $\mathbb{P}(A)\mathbb{P}(B) > 0$
- Two people take turns trying to sink a basketball into a net.
  - Person 1 succeeds with probability $1/3$
  - Person 2 succeeds with probability $1/4$

  What is the probability that person 1 succeeds before person 2?
  Answer: $2/3$

Summary

- The sample space Ω is the set of possible outcomes of an "experiment"
- Points $\omega \in \Omega$ are called realizations
- Events are subsets of Ω
- Properties (axioms) of probability:
  - $0 \leq \mathbb{P}(A) \leq 1$ (Events range from never happening to always happening)
  - $\mathbb{P}(\Omega) = 1$ (Something must happen)
  - $\mathbb{P}(\emptyset) = 0$ (Nothing never happens)
  - $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$ (A must either happen or not-happen)
  - $\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$
- A and B are independent if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$
- Independence is sometimes assumed and sometimes derived.
- Disjoint events with positive probability are not independent.

Math 408 - Mathematical Statistics

Lecture 2. Conditional Probability

January 18, 2013

Agenda

- Motivation and Definition
- Properties of Conditional Probabilities
- Law of Total Probability
- Bayes' Theorem
- Examples
  - False Positive Paradox
  - Monty Hall Problem
- Summary

Motivation and Definition

Recall (see Lecture 1) that the sample space is the set of all possible outcomes of an experiment. Suppose we are interested only in part of the sample space, the part where we know some event – call it A – has happened, and we want to know how likely it is that various other events $(B, C, D \ldots)$ have also happened.

What we want is the **conditional probability** of B given A.

Definition

If $\mathbb{P}(A) > 0$, then the conditional probability of B given A is

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}}$$

Useful Interpretation:

Think of $\mathbb{P}(B|A)$ as the

fraction of times B occurs among those in which A occurs

Properties of Conditional Probabilities

Here are some facts about conditional probabilities:

1. For any fixed A such that $\mathbb{P}(A) > 0$, $\mathbb{P}(\cdot|A)$ is a probability, i.e. it satisfies the rules of probability:
   - $0 \leq \mathbb{P}(B|A) \leq 1$
   - $\mathbb{P}(\Omega|A) = 1$
   - $\mathbb{P}(\emptyset|A) = 0$
   - $\mathbb{P}(B|A) + \mathbb{P}(\bar{B}|A) = 1$
   - $\mathbb{P}(B + C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}(BC|A)$
2. Important: The rules of probability apply to events on the left of the bar.
3. In general
$$\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$$
   Example: the probability of spots given you have measles is 1 but the probability that you have measles given that you have spots is not 1.
4. What if A and B are independent? Then
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$$
   Thus, another interpretation of independence is that knowing A does not change the probability of B.

Law of Total Probability

From the definition of conditional probability we can write

$$\mathbb{P}(AB) = \mathbb{P}(B|A)\mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B)$$

Often these formulae give us a convenient way to compute $\mathbb{P}(AB)$ when A and B are not independent.

A useful tool for computing probabilities is the following law.

Law of Total Probability

Let $A_1, \ldots, A_n$ be a partition of Ω, i.e.

- $\bigcup_{i=1}^n A_i = \Omega$ *($A_1, \ldots, A_k$ are jointly exhaustive events)*
- $A_i \bigcap A_j = \emptyset$ *for* $i \neq j$ *($A_1, \ldots, A_k$ are mutually exclusive events)*
- $\mathbb{P}(A_i) > 0$

Then for any event B

$$\boxed{\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Bayes' Theorem

Conditional probabilities can be inverted. That is,

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}}$$

This relationship is called **Bayes' Rule** after Thomas Bayes (1702-1761) who did not discover it (in this form, Bayes' Rule was proved by Laplace).

Example: False Positive Paradox

Problem

Suppose a rare disease infects one out of every 1000 *people in a population. And suppose that there is a good, but not perfect, test for this disease: if a person has the disease, the test comes back positive* 99% *of the time. One the other hand, the test also produces some false positives. About* 2% *of uninfected patients also test positive. Suppose you just tested positive. What are your chances of having the disease?*

Answer: the chances of having the disease is less than 5% !

Important Conclusion: When dealing with conditional probabilities:

don't trust your intuition, do computations!

Monty Hall problem

Problem

Suppose you are on a game show, and you are given the choice of three doors. A prize is placed at random between one of three doors. You pick a door, say door 1 (but the door is not opened), and the host, who knows what's behind the doors, opens another door which is empty. He then gives you the opportunity to keep your door 1 or switch to the other unopened door. Should you stay or switch?

Answer: You should switch!

Summary

- If $\mathbb{P}(A) > 0$, then

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}}$$

- $\mathbb{P}(\cdot|A)$ satisfies the axioms of probability for fixed A. In general $\mathbb{P}(A|\cdot)$ does not satisfy the axioms of probability for fixed A.
- In general, $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$
- A and B are independent if and only if $\mathbb{P}(B|A) = \mathbb{P}(B)$
- Law of Total Probability: If $A_1, \ldots, A_n$ is a partition of Ω, then for any $B \subset \Omega$

$$\boxed{\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

- Bayes' Theorem

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}}$$

Math 408 - Mathematical Statistics

Lecture 3. Discrete Random Variables

January 23, 2013

Agenda

- Random Variable: Motivation and Definition
- Cumulative Distribution Functions
- Properties of CDFs
- Discrete Random Variables
- Important Examples
  - The Point Mass Distribution
  - The Discrete Uniform Distribution
  - The Bernoulli Distribution
  - The Binomial Distribution
  - The Geometric Distribution
  - The Poisson Distribution
- Summary

Motivation and Definition

Statistics is concerned with data.

Question: How do we link sample spaces and events to data?

Answer: The link is provided by the concept of a **random variable**.

Definition

A random variable is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $x = X(\omega)$ to each realization $\omega \in \Omega$.

Remark: Technically, a random variable must be a measurable function.

Example: Flip a coin 10 times. Let $X(\omega)$ be the number of heads in the sequence. For example, if $\omega = HHTHTTTHTH$, then $X(\omega) = 5$.

Given a random variable X and a set $A \subset \mathbb{R}$, define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and let

$$\boxed{\begin{aligned}\mathbb{P}(X \in A) =& \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \\ \mathbb{P}(X = x) =& \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})\end{aligned}}$$

The Cumulative Distribution Function

Definition

The cumulative distribution function (CDF) $F_X : \mathbb{R} \to [0, 1]$ is defined by

$$\boxed{F_X(x) = \mathbb{P}(X \leq x)}$$

Example: Flip a fair coin twice and let X be the number of heads.
Find the CDF of X

Question: Why do we bother to define CDF?

Answer: CDF effectively contains all the information about the random variable

Theorem

Let X have CDF F and Y have CDF G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$. In words, the CDF completely determines the distribution of a random variable.

Properties of CDFs

Question: Given a function $F(x)$, can we find a random variable X such that $F(x)$ is the CDF of X, $F_X(x) = F(x)$?

Theorem

A function $F : \mathbb{R} \to [0, 1]$ is a CDF for some random variable if and only if it satisfies the following three conditions:

1. *F is non-decreasing:*

$$x_1 < x_2 \quad \Rightarrow \quad F(x_1) \leq F(x_2)$$

2. *F is normalized:*

$$\lim_{x \to -\infty} F(x) = 0 \quad \textit{and} \quad \lim_{x \to +\infty} F(x) = 1$$

3. *F is right-continuous:*

$$\lim_{y \to x+0} F(y) = F(x)$$

Discrete Random Variables

Definition

X is **discrete** if it takes countable many values $\{x_1, x_2, \dots\}$.
We define the **probability mass function** (PMF) for X by

$$\boxed{f_X(x) = \mathbb{P}(X = x)}$$

Example: Flip a fair coin twice and let X be the number of heads.
Find the probability mass function of X.

The CDF of X is related to the PMF f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

The PMF f_X is related to the CDF F_X by

$$f_X(x) = F_X(x) - F_X(x^-) = F_X(x) - \lim_{y \to x-0} F(y)$$

Important Examples

- **The Point Mass Distribution**
  X has a point mass distribution at a, denoted $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$.
  In this case
  $$F(x) = \begin{cases} 0, & x < a; \\ 1, & x \geq a. \end{cases}$$
  and
  $$f(x) = \begin{cases} 1, & x = a; \\ 0, & x \neq a. \end{cases}$$

- **The Discrete Uniform Distribution**
  Let $n > 1$ be a given integer. Suppose that X has probability mass function given by
  $$f(x) = \begin{cases} 1/n, & \text{for } x = 1, \dots, n; \\ 0, & \text{otherwise.} \end{cases}$$
  We say that X has a uniform distribution on $1, \dots, n$.

Important Examples

- **The Bernoulli Distribution**
  Let X represents a coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution, denoted $X \sim \mathrm{Bernoulli}(p)$. The probability mass function is

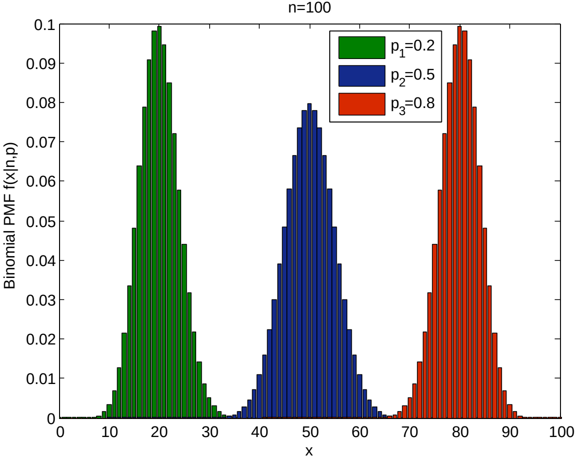
$$f(x|p) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

- **The Binomial Distribution**
  Suppose we have a coin which falls heads with probability p for some $p \in [0, 1]$. Flip the coin n times and let X be the number of heads. Assume that the tosses are independent. The probability mass function of X is then

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

  A random variable with this mass function is called a Binomial random variable and we write $X \sim \mathrm{Bin}(n, p)$.
  Remark: X is a random variable, x denotes a particular value of the random variable, n and p are parameters, that is, fixed real numbers. The parameter p is usually unknown and must be estimated from data.

Binomial Distribution $\mathrm{Bin}(n, p)$



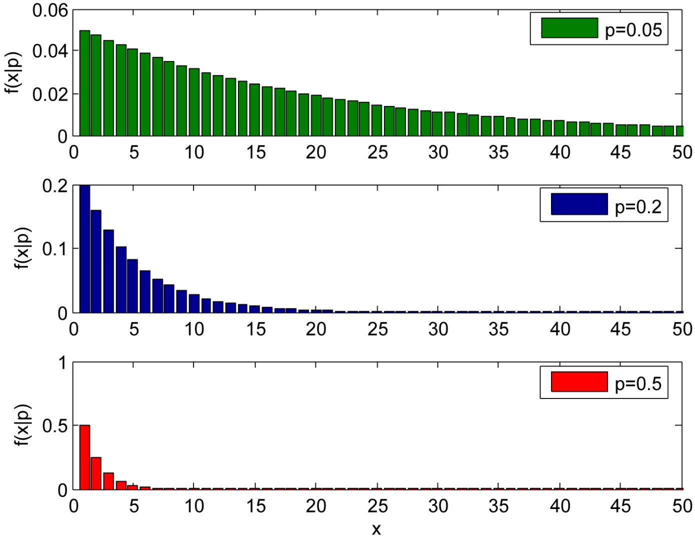
Important Examples

- **The Geometric Distribution**
  X has a geometric distribution with parameter $p \in (0,1)$, denoted $X \sim \mathrm{Geom}(p)$, if

$$f(x|p) = p(1-p)^{x-1}, \quad x = 1, 2, 3 \ldots$$

  Think of X as the number of flips needed until the first heads when flipping a coin. Geometric distribution is used for modeling the number of trials until the first success.

Geometric Distribution $\mathrm{Geom}(p)$



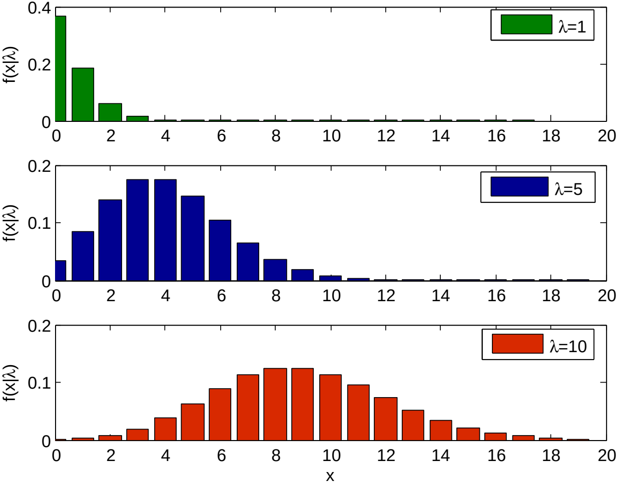
Important Examples

- **The Poisson Distribution**
  X has a Poisson distribution with parameter λ,
  denoted $X \sim \mathrm{Poisson}(\lambda)$ if

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

  The Poisson distribution is often used as a model for counts of rare events like traffic accidents. $f(x|\lambda)$ expresses the probability of a given number of events x occurring in a fixed interval of time if these events occur with a known average rate λ and independently of the time since the last event.

Poisson Distribution $\text{Poisson}(\lambda)$



Summary

- A random variable is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $x = X(\omega)$ to each realization $\omega \in \Omega$.
- The cumulative distribution function (CDF) is defined by

$$\boxed{F_X(x) = \mathbb{P}(X \leq x)}$$

  - CDF completely determines the distribution of a random variable
  - CDF is non-decreasing, normalized, and right-continuous
- Random variable X is discrete if it takes countable many values $\{x_1, x_2, \ldots\}$.
- The probability mass function (PMF) of X is

$$\boxed{f_X(x) = \mathbb{P}(X = x)}$$

- Relationships between CDF and PMF:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

$$f_X(x) = F_X(x) - F_X(x^-)$$

Math 408 - Mathematical Statistics

Lecture 4. Continuous Random Variables and Transformations of Random Variables

January 25, 2013

Agenda

- Definition
- Important Examples
  - Uniform Distribution
  - Normal (Gaussian) Distribution
  - Exponential Distribution
  - Gamma Distribution
  - Beta Distribution
- Transformation of Random Variables
  - Discrete Case
  - Continuous Case
- Summary

Definition

Recall that a random variable is a (deterministic) map $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each (random) realization $\omega \in \Omega$.

Definition

A random variable is **continuous** if there exists a function f_X such that

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$, and
- For every $a \leq b$

$$\boxed{P(a < X \leq b) = \int_a^b f_X(x)dx}$$

- The function $f_X(x)$ is called the probability density function (PDF)
- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)dt} \qquad \boxed{f_X(x) = F_X'(x)}$$

Important Remarks

- If X is continuous then $\mathbb{P}(X = x) = 0$ for every x.
- Don't think of $f_X(x)$ as $\mathbb{P}(X = x)$. This is only true for discrete random variables.
- For continuous random variables, we get probabilities by integrating.
- A PDF can be bigger than 1 (unlike PMF!). For example:

$$f_X(x) = \begin{cases} 10, & x \in [0, 0.1] \\ 0, & x \notin [0, 0.1] \end{cases}$$

- Can a PDF be unbounded?
  Yes, of course! For instance

$$f_X(x) = \begin{cases} \frac{2}{3}x^{-1/3}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Important Examples

- **The Uniform Distribution**
  X has a uniform distribution on $[a, b]$, denoted $X \sim U[a, b]$, if

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

- **Normal (Gaussian) Distribution**
  X has a Normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if

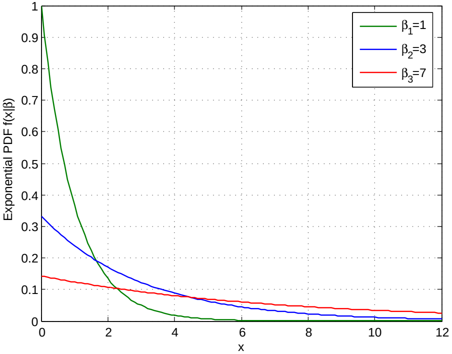
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

  - Many phenomena in nature have approximately Normal distribution.
  - Distribution of a sum of random variables can be approximated by a Normal distribution (central limit theorem)

Important Examples

- **Exponential Distribution**
  X has an Exponential distribution with parameter $\beta > 0$, $X \sim \mathrm{Exp}(\beta)$, if

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0$$



The exponential distribution is used to model the life times of electronic components and the waiting times between rare events. β is a survival parameter: the expected duration of survival of the system is β units of time.

Important Examples

- **Gamma Distribution**
  X has a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, $X \sim \mathrm{Gamma}(\alpha, \beta)$, if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

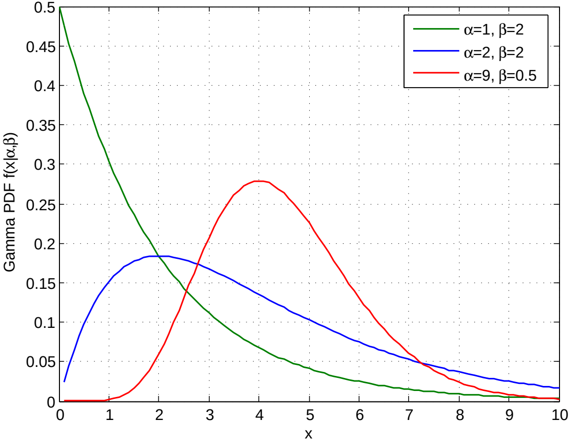
  - $\Gamma(\alpha)$ is the Gamma function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

  - The Gamma distribution is frequently used to model waiting times.
  - Exponential distribution is a special case of the Gamma distribution:

$$\mathrm{Gamma}(1, \beta) = \mathrm{Exp}(\beta)$$

Gamma Distribution



Important Examples

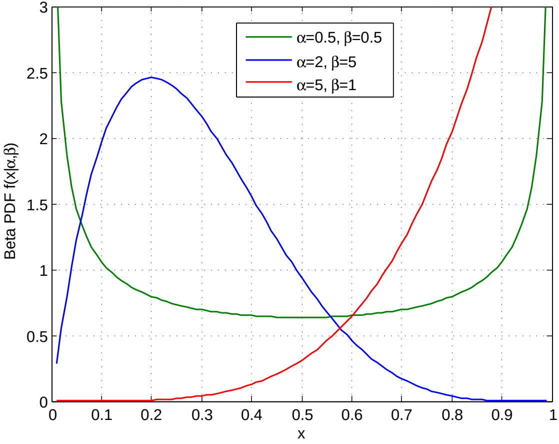
- **Beta Distribution**
  X has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, $X \sim \mathrm{Beta}(\alpha, \beta)$, if

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

  - The beta distribution is often used for modeling of proportions.
  - The beta distribution has an important application in the theory of order statistics. A basic result is that the distribution of the k^{th} largest $X_{(k)}$ of a sample of size n from a uniform distribution $X_1, \ldots, X_n \sim U(0,1)$ has a beta distribution:

$$X_{(k)} \sim \mathrm{Beta}(k, n-k+1)$$

Beta Distribution



Transformation of Random Variables

Suppose that X is a random variable with PDF/PMF (continuous/discrete) f_X and CDF F_X. Let $Y = r(X)$ be a function of X, for example, $Y = X^2$, $Y = e^X$.

Q: How to compute the PDF/PMF and CDF of Y?

In the discrete case, the answer is easy:

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(r(X) = y) = \mathbb{P}(\{x : r(x) = y\}) = \sum_{x_i : r(x_i) = y} f_X(x_i)$$

Example:

- $X \in \{-1, 0, 1\}$
- $\mathbb{P}(X = -1) = 1/4, \quad \mathbb{P}(X = 0) = 1/2, \quad \mathbb{P}(X = 1) = 1/4$
- $Y = X^2$
- Find PMF f_Y

Answer: $f_Y(0) = 1/2$ and $f_Y(1) = 1/2$.

Transformation of Random Variables: Continuous Case

The continuous case is harder.
These are the steps for finding the PDF f_Y:

1. For each y, let $A_y = \{x : r(x) \leq y\}$
2. Find the CDF $F_Y(y)$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \in A_y) = \int_{A_y} f_X(x)dx$$

3. The PDF is then $f_Y(y) = F'_Y(y)$

Example: Let $X \sim \mathrm{Exp}(1)$, and $Y = \ln X$. Find $f_Y(y)$.

Answer: $f_Y(y) = e^y e^{-e^y}$, $y \in \mathbb{R}$

Important Fact: When r is strictly monotonic, then r has an inverse $s = r^{-1}$ and

$$\boxed{f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|}$$

Summary

- A random variable is continuous if there exists a function f_X, called probability density function such that
  - $f_X(x) \geq 0$ for all x
  - $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
  - For every $a \leq b$

$$\boxed{P(a < X \leq b) = \int_a^b f_X(x)dx}$$

- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)dt} \qquad \boxed{f_X(x) = F'_X(x)}$$

- Important Examples: Uniform Distribution, Normal Distribution, Exponential Distribution, Gamma Distribution, Beta Distribution
- If $Y = r(X)$ and r is strictly monotonic, then

$$\boxed{f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|} \qquad s = r^{-1}$$

Math 408 - Mathematical Statistics

Lecture 5. Joint Distributions

January 28, 2013

Agenda

- Bivariate Distributions
- Marginal Distributions
- Independent Random Variables
- Conditional Distributions
- Transformation of Several Random Variables
- Summary

Bivariate Distributions

- Discrete Case

Definition

Given a pair of discrete random variables X and Y, their **joint PMF** is defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$$

- Continuous Case

Definition

A function $f_{X,Y}(x,y)$ is called the **joint PDF** of continuous random variables X and Y if

- $f_{X,Y}(x,y) \geq 0, \quad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$
- For any set $A \subset \mathbb{R} \times \mathbb{R}$

$$\mathbb{P}((X,Y) \in A) = \int\int_A f_{X,Y}(x,y)dxdy$$

The **joint CDF** of X and Y is defined as $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$

Marginal Distributions

- Discrete Case
  If X and Y have joint PMF $f_{X,Y}$, then the **marginal PMF** of X is

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

Similarly, the **marginal PMF** of Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f_{X,Y}(x, y)$$

- Continuous Case
  If X and Y have joint PDF $f_{X,Y}$, then the **marginal PDFs** of X and Y are

$$f_X(x) = \int f_{X,Y}(x, y)dy \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x, y)dx$$

Examples

- Suppose that the PMF f_{XY} is given in the following table:

	$Y=0$	$Y=1$
$X=0$	$1/10$	$2/10$
$X=1$	$3/10$	$4/10$

Find the marginal PMF of X.
<u>Answer:</u> $f_X(0)=3/10$, $f_X(1)=7/10$

- Suppose that

$$f_{X,Y}(x,y)=e^{-(x+y)}, \quad x,y\geq 0$$

Find the marginal PDF of X.
<u>Answer:</u> $f_X(x)=e^{-x}$, $x\geq 0$

Independent Random Variables

Definition

Two random variables X and Y are **independent** if, for every A and B

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

In principle, to check whether X and Y are independent, we need to check the above equation for all subsets A and B. Fortunately, we have the following result:

Theorem

Let X and Y have joint PDF/PMF $f_{X,Y}$. Then X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Example: Suppose that X and Y are independent and both have the same density

$$f(x) = \begin{cases} 2x, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$$

Find $\mathbb{P}(X + Y \leq 1)$. Answer: $1/6$

Conditional Distributions

- Discrete Case

  If X and Y are discrete, then we can compute the conditional probability of the event $\{X = x\}$ given that we have observed $\{Y = y\}$:

  $$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

  This leads to the following definition of the **conditional PMF**:

  $$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Continuous Case

  For continuous random variables, the **conditional PDF** is

  $$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

  Then,

  $$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$$

Example

- Suppose that $X \sim U(0,1)$. After obtaining a value x of X, we generate $Y|X = x \sim U(x,1)$. What is the marginal distribution of Y?
  Answer:
  $$f_Y(y) = -\ln(1-y) \quad y \in (0,1)$$

Transformation of Several Random Variables

In some cases we are interested in transformation of several random variables. For example, if X and Y are given random variables, we might want to know the distribution of X/Y, $X+Y$, $\max\{X, Y\}$, etc.

Let $Z = r(X, Y)$. The steps for finding f_Z are the following:

1. For each z, find the set $A_z = \{(x, y) : r(x, y) \leq z\}$
2. Find the CDF

$$F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(r(X, Y) \leq z) = \mathbb{P}((X, Y) \in A_z) = \int\int_{A_z} f_{X,Y}(x, y)dxdy$$

3. Then PDF $f_Z(z) = F'_Z(z)$

Example: Let $X, Y \sim U[0, 1]$ be independent.
Find the density of $Z = X + Y$.

Answer:

$$f_Z(z) = \begin{cases} z, & 0 \leq z \leq 1 \\ 2 - z, & 1 < z \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Summary

- Joint Distributions:
  - Discrete case: $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$
  - Continuous case: $\mathbb{P}((X,Y) \in A) = \int\int_A f_{X,Y}(x,y)dxdy$
- Marginal Distributions
  - Discrete case: $f_X(x) = \sum_y f_{X,Y}(x,y)$
  - Continuous case: $f_X(x) = \int f_{X,Y}(x,y)dy$
- X and Y are independent if, for every A and B

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

  - X and Y are independent if and only if $\boxed{f_{X,Y}(x,y) = f_X(x)f_Y(y)}$
- Conditional PDF/PMF:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Algorithm for finding distribution of $Z = r(X,Y)$

Math 408 - Mathematical Statistics

Lecture 6. Expectation, Variance, Covariance, and Correlation

January 30, 2013

Expectation of a Random Variable

The expectation (or mean) of a random variable X is the average value of X. The formal definition is as follows.

Definition

The **expected value**, or **mean**, or **first moment** of X is

$$\boxed{\mu_X \equiv \mathbb{E}[X] = \begin{cases} \sum_x x f_X(x), & \text{if } X \text{ is discrete} \\ \int x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}}$$

assuming that the sum (or integral) is well-defined.

Remarks:

- The expectation is a one-number summary of the distribution.
- Think of $\mathbb{E}[X]$ as the average value you would obtain if you computed the numerical average $\frac{1}{n}\sum_{i=1}^{n} X_i$ of a large number of i.i.d. draws $X_1, \ldots, X_n$. The fact that

$$\mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a theorem called the law of large numbers.

Examples

- Let $X \sim \mathrm{Bernoulli}(p)$. Find $\mathbb{E}[X]$.
  Answer: $\mathbb{E}[X] = p$
- Let $X \sim U(-1, 3)$. Find $\mathbb{E}[X]$.
  Answer: $\mathbb{E}[X] = 1$

Let $Y = r(X)$. How do we compute $\mathbb{E}[Y]$? There are two ways:

- Find $f_Y(y)$ (Lecture 4) and then compute $\mathbb{E}[Y] = \int y f_Y(y) dy$.
- An easier way:

$$\boxed{\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) f_X(x) dx}$$

Example: Take a stick of unit length and break it at random. Let Y be the length of the longer piece. What is the mean of Y?
Answer: $\mathbb{E}[Y] = \frac{3}{4}$
Functions of several variables are handled in a similar way: if $Z = r(X, Y)$, then

$$\boxed{\mathbb{E}[Z] = \mathbb{E}[r(X, Y)] = \int \int r(x, y) f_{X,Y}(x, y) dx dy}$$

Properties of Expectations

- If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\boxed{\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]}$$

  - Let $X \sim \mathrm{Bin}(n, p)$. Find $\mathbb{E}[X]$.
  - Answer: $\mathbb{E}[X] = np$

- Let $X_1, \ldots, X_n$ be independent random variables. Then,

$$\boxed{\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]}$$

Remark: Note the the summation rule does not require independence but the multiplication rule does.

Variance and Its Properties

The variance measures the "spread" of a distribution.

Definition

Let X be a random variance with mean μ_X.
The **variance** of X, denoted $\mathbb{V}[X]$ or σ_X^2, is defined by

$$\boxed{\sigma_X^2 \equiv \mathbb{V}[X] = \mathbb{E}[(X-\mu_X)^2] = \begin{cases} \sum_x (x-\mu_X)^2 f_X(x), & \text{if } X \text{ is discrete} \\ \int (x-\mu_X)^2 f_X(x)dx, & \text{if } X \text{ is continuous} \end{cases}}$$

The **standard deviation** is $\sigma_X = \sqrt{\mathbb{V}[X]}$

Important Properties of $\mathbb{V}[X]$:

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mu_X^2$
- If a and b are constants, then $\mathbb{V}[aX+b] = a^2\mathbb{V}[X]$
- If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i]$$

Covariance and Correlation

Example: Let $X \sim \mathrm{Bin}(n, p)$. Find $\mathbb{V}[X]$.
Answer: $\mathbb{E}[X] = np(1-p)$

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the **covariance** between X and Y by

$$\boxed{\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}$$

and the **correlation** by

$$\boxed{\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}}$$

Properties of Covariance and Correlation

- The covariance satisfies (useful in computations):

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1$$

- If $Y = aX + b$ for some constants a and b, then

$$\rho(X, Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0 \end{cases}$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0$. The converse is not true.
- For random variables $X_1, \ldots, X_n$

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

Expectation and Variance of Important Random Variables

Distribution	**Mean**	**Variance**
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Bin(n,p)	p	$np(1-p)$
Geom(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
Exp(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$

Summary

- The expected value of X is

$$\boxed{\mu_X \equiv \mathbb{E}[X] = \begin{cases} \sum_x x f_X(x), & \text{if } X \text{ is discrete} \\ \int x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}}$$

  - If $Y = r(X)$, then $\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) f_X(x) dx$
  - If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then $\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i \mathbb{E}[X_i]$
  - If $X_1, \ldots, X_n$ are independent random variables, then $\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i]$

- The variance of X is

$$\boxed{\sigma_X^2 \equiv \mathbb{V}[X] = \mathbb{E}[(X - \mu_X)^2]}$$

  - $\mathbb{V}[X] = \mathbb{E}[X^2] - \mu_X^2$
  - If a and b are constants, then $\mathbb{V}[aX + b] = a^2 \mathbb{V}[X]$
  - If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then $\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i]$

Summary

- Covariance and correlation between X and Y are

$$\boxed{\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]} \qquad \boxed{\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}}$$

  - $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
  - $-1 \leq \rho(X,Y) \leq 1$
  - If $Y = aX + b$ then $\rho(X,Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0 \end{cases}$
  - If X and Y are independent, then $\mathrm{Cov}(X,Y) = \rho(X,Y) = 0$.
  - $\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$

Math 408 - Mathematical Statistics

Lecture 7. Conditional Expectation and Conditional Variance

February 1, 2013

Definition

Suppose that X and Y are random variables.

Q: What is the mean of X among those times when $Y = y$?

A: It is the mean of X as before, but instead of $f_X(x)$ we use $f_{X|Y}(x|y)$.

Definition

The **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum_x x f_{X|Y}(x|y), & \text{discrete case;} \\ \int x f_{X|Y}(x|y)dx, & \text{continuous case.} \end{cases}$$

If $Z = r(X, Y)$ is a new random variable, then

$$\mathbb{E}[Z|Y = y] = \begin{cases} \sum_x r(x,y) f_{X|Y}(x|y), & \text{discrete case;} \\ \int r(x,y) f_{X|Y}(x|y)dx, & \text{continuous case.} \end{cases}$$

Important Remark:

- $\mathbb{E}[X]$ is a number
- $\mathbb{E}[X|Y = y]$ is a function of y

Conditional Expectation

Question: What is $\mathbb{E}[X|Y=y]$ before we observe the value y of Y?

Answer: Before we observe Y, we don't know the value of $\mathbb{E}[X|Y=y]$, it is uncertain, so it is a random variable which we denote $\mathbb{E}[X|Y]$.

$\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y=y]$ when $Y=y$.

Example 1:
Suppose we draw

$$X \sim U(0,1)$$

After we observe $X=x$, we draw

$$Y|X=x \sim U(x,1)$$

Find $\mathbb{E}[Y|X=x]$.
Answer:

$$\mathbb{E}[Y|X=x] = \frac{x+1}{2}, \qquad \text{as intuitively expected}$$

Note that $\mathbb{E}[Y|X] = \frac{X+1}{2}$ is a random variable whose value is the number $\mathbb{E}[Y|X=x] = \frac{x+1}{2}$ once $X=x$ is observed.

The Rule of Iterated Expectations

Theorem

For random variables X and Y, assuming the expectations exist, we have

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

More generally, for any function $r(x, y)$ we have

$$\mathbb{E}[\mathbb{E}[r(X,Y)|X]] = \mathbb{E}[r(X,Y)] \quad \text{and} \quad \mathbb{E}[\mathbb{E}[r(X,Y)|Y]] = \mathbb{E}[r(X,Y)]$$

Example 2: Compute $\mathbb{E}[Y]$ in Example 1.

Answer:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}\left[\frac{X+1}{2}\right] = \frac{1/2+1}{2} = \frac{3}{4}$$

Conditional Variance

Recall, that "unconditional" variance of random variable Y is

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

Therefore, it is natural to define **conditional variance** of Y given that $X = x$ as follows (replace all expectations by conditional expectations):

$$\mathbb{V}[Y|X = x] = \mathbb{E}[(Y - \mathbb{E}[Y|X = x])^2|X = x]$$

Denote $\mathbb{E}[Y|X = x]$ by $\mu_Y(x)$. Then

$$\mathbb{V}[Y|X = x] = \int (y - \mu_Y(x))^2 f_{Y|X}(y|x)dy$$

- $\mathbb{V}[Y]$ is a number, $\mathbb{V}[Y|X = x]$ is a function of x

Theorem

For random variables X and Y

$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$$

Example: Statistical Analysis of a Disease

- Draw a state at random from the US.
- Let Q be the proportion of people in that state with a certain disease. Q is a random variable since it varies from state to state, and state is picked at random.
  - Suppose that Q has a uniform distribution on $(0,1)$, $Q \sim U(0,1)$.
  - This assumption is natural if we don't have any information about the disease.
- Draw n people at random from the state, and let X be the number of those people who have the disease.
  - Given $Q = q$, it is natural to model X as a Binomial variable, $X|Q = q \quad \sim \quad \mathrm{Bin}(n, q)$.

Problem: Find $\mathbb{E}[X]$ and $\mathbb{V}[X]$

Answer:

$$\mathbb{E}[X] = \frac{n}{2}$$

$$\mathbb{V}[X] = \frac{n}{6} + \frac{n^2}{12}$$

Summary

- The conditional expectation of X given $Y=y$ is

$$\mathbb{E}[X|Y=y]=\begin{cases} \sum_x x f_{X|Y}(x|y), & \text{discrete case;} \\ \int x f_{X|Y}(x|y)dx, & \text{continuous case.} \end{cases}$$

  - $\mathbb{E}[X]$ is a number
  - $\mathbb{E}[X|Y=y]$ is a function of y
  - $\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y=y]$ when $Y=y$

- The Rule of Iterated Expectations

$$\mathbb{E}\mathbb{E}[Y|X]=\mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}\mathbb{E}[X|Y]=\mathbb{E}[X]$$

- The conditional variance of X given $Y=y$ is

$$\mathbb{V}[X|Y=y]=\mathbb{E}[(X-\mathbb{E}[X|Y=y])^2|Y=y]$$

  - $\mathbb{V}[X]$ is a number
  - $\mathbb{V}[X|Y=y]$ is a function of y
  - $\mathbb{V}[X|Y]$ is the random variable whose value is $\mathbb{V}[X|Y=y]$ when $Y=y$

- For random variables X and Y

$$\mathbb{V}[X]=\mathbb{E}\mathbb{V}[X|Y]+\mathbb{V}\mathbb{E}[X|Y]$$

Math 408 - Mathematical Statistics

Lecture 8. Inequalities

February 4, 2013

Agenda

- Markov Inequality
- Chebyshev Inequality
- Hoeffding Inequality
- Cauchy-Schwarz Inequality
- Jensen Inequality
- Summary

Markov Inequality

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will be used in the large sample theory (next two Lectures) which is extremely important for statistical inference.

Markov Inequality

Let X be a non-negative random variable and suppose that $\mathbb{E}[X]$ exists. Then for any $a > 0$

$$\boxed{\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}}$$

Remark:

- This result says that the probability that X is much bigger than $\mathbb{E}[X]$ is small: Let

$$a = k\mathbb{E}[X]$$

Then

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Chebyshev Inequality

Chebyshev Inequality

Let X be a random variable with mean μ and variance σ^2. Then for any $a > 0$

$$\boxed{\mathbb{P}(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}}$$

Remarks:

- This result says that if σ^2 is small, then there is a high probability that X will not deviate much from μ.
- If $a = k\sigma$, then

$$\mathbb{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- If $Z = \frac{X-\mu}{\sigma}$, then

$$\mathbb{P}(|Z| \geq a) \leq \frac{1}{a^2}$$

Example

Suppose we test a prediction method on a set of n new test cases. Let

$$X_i = \begin{cases} 1, & \text{if the predictor is wrong;} \\ 0, & \text{if the predictor is right.} \end{cases}$$

Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the observed error rate. Let p be the true error rate. We hope that $\overline{X}_n \approx p$.

Question: Estimate the probability $\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon)$

Answer:

$$\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

Hoeffding Inequality

Hoeffding inequality is similar in spirit to Chebyshev inequality but it is sharper. This is how it looks in a special case for Bernoulli random variables:

Hoeffding Inequality

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Then for any* $\varepsilon > 0$

$$\boxed{\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}}$$

Remark: Hoeffding inequality gives us a simple way to create a confidence interval for a binomial parameter p.

Definition

A $100(1-\alpha)\%$ confidence interval for a parameter p is an interval calculated from the sample $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, which contains p with probability $1-\alpha$.

Example: Construct a $100(1-\alpha)\%$ confidence interval for p using Hoeffding inequality.

Answer: $\overline{X}_n \pm \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}$

Cauchy-Schwarz and Jensen Inequalities

These are two inequalities on expected values that are often useful.

Cauchy-Schwarz Inequality

If X and Y have finite variances, then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Jensen Inequality

- *If g is convex (x^2, e^x, etc), then*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

- *If g is concave ($-x^2$, $\log x$, etc), then*

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

Examples: $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, $\mathbb{E}(1/X) \geq 1/\mathbb{E}[X]$, $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$.

Summary

- Markov inequality: If X is a non-negative random variable, then for any $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

- Hoeffding inequality: Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$, then for any $\varepsilon > 0$

$$\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

- Cauchy-Schwarz inequality: If X and Y have finite variances, then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

- Jensen Inequality:
  - If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
  - If g is concave, then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$

Math 408 - Mathematical Statistics

Lecture 9-10. Tricks with Random Variables: The Law of Large Numbers & The Central Limit Theorem

February 6-8, 2013

Agenda

- Large Sample Theory
- Types of Convergence
  - Convergence in Probability
  - Convergence in Distribution
- The Law of Large Numbers
  - The Monte Carlo Method
- The Central Limit Theorem
  - Multivariate version
- Summary

Large Sample Theory

The most important aspect of probability theory concerns the behavior of sequences of random variables. This part of probability is called large sample theory or limit theory or asymptotic theory. This theory is extremely important for statistical inference.

The basic question is this:

What can we say about the limiting behavior of a sequence of random variables?

$$X_1, X_2, X_3 \dots$$

In the statistical context: What happens as we gather more and more data?

In Calculus, we say that a sequence of real numbers $x_1, x_2, \dots$ converges to a limit x if, for every $\epsilon > 0$, we can find N such that $|x_n - x| < \epsilon$ for all $n > N$.

In Probability, convergence is more subtle.

Going back to calculus, suppose that $x_n = 1/n$. Then trivially, $\lim_{n\to\infty} x_n = 0$. Consider a probabilistic version of this example: suppose that $X_1, X_2, \dots$ are independent and $X_n \sim \mathcal{N}(0, 1/n)$. Intuitively, X_n is very concentrated around 0 for large n, and we are tempted to say that X_n "converges" to zero. However, $\mathbb{P}(X_n = 0) = 0$ for all n!

Types of Convergence

There are two main types of convergence:
convergence in probability and convergence in distribution

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

- X_n **converges** to X **in probability**, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$
- X_n **converges** to X **in distribution**, written $X_n \xrightarrow{\mathcal{D}} X$, if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for all x for which F is continuous.

Relationships Between the Types of Convergence

Example: Let $X_n \sim \mathcal{N}(0, 1/n)$. Then

- $X_n \xrightarrow{\mathbb{P}} 0$
- $X_n \xrightarrow{\mathcal{D}} 0$

Question: Is there any relationship between $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathcal{D}}$?

Answer: Yes:

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \quad \text{implies that} \quad X_n \xrightarrow{\mathcal{D}} X}$$

Important Remark: The reverse implication does not hold: convergence in distribution does not imply convergence in probability.

Example: Let $X \sim \mathcal{N}(0, 1)$ and let $X_n = -X$ for all n. Then

- $X_n \xrightarrow{\mathcal{D}} X$
- $X_n \overset{\mathbb{P}}{\nrightarrow} X$

The Law of Large Numbers

The law of large numbers is one of the main achievements in probability. This theorem says that the mean of a large sample is close to the mean of the distribution.

The Law of Large Numbers

Let $X_1, X_2, \dots$ be an i.i.d. sample and let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathbb{V}[X_1] < \infty$. Then

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu}$$

Useful Interpretation:

The distribution of $\overline{X}_n$ becomes more concentrated around μ as n gets larger.

Example: Let $X_i \sim \mathrm{Bernoulli}(p)$. The fraction of heads after n tosses is $\overline{X}_n$.

According to the LLN, $\overline{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_i] = p$. It means that, when n is large, the distribution of $\overline{X}_n$ is tightly concentrated around p.

Q: How large should n be so that $\mathbb{P}(|\overline{X}_n - p| < \epsilon) \geq 1 - \alpha$?

Answer: $n \geq \frac{p(1-p)}{\alpha\epsilon^2}$

The Monte Carlo Method

Suppose we want to calculate

$$I(f) = \int_0^1 f(x)dx$$

where the integration cannot be done by elementary means.

The Monte Carlo method works as follows:

1. Generate independent uniform random variables on [0,1], $X_1, \ldots, X_n \sim U[0,1]$
2. Compute $Y_1 = f(X_1), \ldots, Y_n = f(X_n)$. Then $Y_1, \ldots, Y_n$ are i.i.d.
3. By the law of large numbers $\overline{Y}_n$ should be close to $\mathbb{E}[Y_1]$. Therefore:

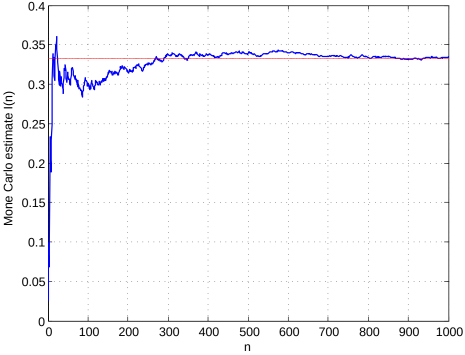
$$\boxed{\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \overline{Y}_n \approx \mathbb{E}[Y_1] = \mathbb{E}[f(X_1)] = \int_0^1 f(x)dx}$$

Monte Carlo method: Example

Suppose we want to compute the following integral:

$$I = \int_0^1 x^2 dx$$

- From calculus: $I = 1/3$
- Using Monte Carlo method: $I(n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, where $X_i \sim U[0,1]$



Accuracy of the Monte Carlo method

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \int_0^1 f(x)dx, \qquad X_1,\ldots,X_n \sim U[0,1]$$

Question: How large should n be to achieve a desired accuracy?

Answer: Let $f:[0,1]\to[0,1]$. To get $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ within ϵ of the true value $I(f)$ with probability at least p, we should choose n so that

$$\boxed{n \geq \frac{1}{\epsilon^2(1-p)}}$$

Thus, the Monte Carlo method tells us how large to take n to get a desired accuracy.

The Central Limit Theorem

Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. The **central limit theorem** (CLT) says that $\overline{X}_n$ has a distribution which is approximately Normal. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

Useful Interpretation:

- Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution.

The Central Limit Theorem

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

There are several forms of notation to denote the fact that the distribution of Z_n is converging to a Normal. They all mean the same thing:

$$Z_n \;\dot{\sim}\; \mathcal{N}(0,1)$$

$$\overline{X}_n \;\dot{\sim}\; \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\overline{X}_n - \mu \;\dot{\sim}\; \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\overline{X}_n - \mu) \;\dot{\sim}\; \mathcal{N}(0, \sigma^2)$$

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \;\dot{\sim}\; \mathcal{N}(0,1)$$

The Central Limit Theorem: Remarks

- The CLT asserts that the CDF of $\overline{X}_n$, suitably normalized to have mean 0 and variance 1, converges to the CDF of $\mathcal{N}(0,1)$.

  Q: Is the corresponding result valid at the level of PDFs and PMFs?

  Broadly speaking the answer is yes, but some condition of smoothness is necessary (generally, $F_n(x) \to F(x)$ does not imply $F_n'(x) \to F'(x)$).
- The CLT does not say anything about the rate of convergence.
- The CLT tells us that in the long run we know what the distribution must be.
  - Even better: it is always the same distribution.
    - Still better: it is one which is remarkably easy to deal with, and for which we have a huge amount of theory.

Historic Remark:

- For the special case of Bernoulli variables with $p = 1/2$, CLT was proved by **de Moivre** around **1733**.
- General values of p were treated later by **Laplace**.
- The first rigorous proof of CLT was discovered by **Lyapunov** around **1901**.

The Central Limit Theorem: Example

- Suppose that the number of errors per computer program has a Poisson distribution with mean $\lambda = 5$. $f(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$
- We get $n = 125$ programs; n is sample size
- Let $X_1, \ldots, X_n$ be the number of errors in the programs, $X_i \sim \mathrm{Poisson}(\lambda)$.
- Estimate probability $\mathbb{P}(\overline{X}_n \leq \lambda + \epsilon)$, where $\epsilon = 0.5$.

Answer:

$$\mathbb{P}(\overline{X}_n \leq \lambda + \epsilon) \approx \Phi\left(\epsilon\sqrt{\frac{n}{\lambda}}\right) = \Phi(2.5) \approx 0.994$$

The Central Limit Theorem: Example

- A tourist in Las Vegas was attracted by a certain gambling game in which
  - the customer stakes 1 dollar on each play
  - a win then pays the customer 2 dollars plus the return of her stake
  - a loss costs her only her stake
- The probability of winning at this game is $p = 1/4$.
- The tourist played this game $n = 240$ times.

Assuming that no near miracles happened,

- about how much poorer was the tourist upon leaving the casino?
  Answer:
$$\mathbb{E}[\text{payoff}] = -\$60$$
- what is the probability that she lost no money?
  Answer:
$$\mathbb{P}[\text{payoff} \geq 0] \approx 0$$

The Central Limit Theorem

The central limit theorem tells us that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \ \dot{\sim}\ \mathcal{N}(0,1)$$

However, in applications, we rarely know σ. We can estimate σ^2 from $X_1, \ldots, X_n$ by sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Question: If we replace σ with S_n is the central limit theorem still true?

Answer: Yes!

Theorem

Assume the same conditions as the CLT. Then,

$$\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

Multivariate Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. random vectors with mean μ and covariance matrix Σ:

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_{1i}] \\ \mathbb{E}[X_{2i}] \\ \vdots \\ \mathbb{E}[X_{ki}] \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \mathbb{V}[X_{1i}] & \mathrm{Cov}(X_{1i}, X_{2i}) & \ldots & \mathrm{Cov}(X_{1i}, X_{ki}) \\ \mathrm{Cov}(X_{2i}, X_{1i}) & \mathbb{V}[X_{2i}] & \ldots & \mathrm{Cov}(X_{2i}, X_{ki}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{ki}, X_{1i}) & \ldots & \mathrm{Cov}(X_{ki}, X_{k-1i}) & \mathbb{V}[X_{ki}] \end{pmatrix}$$

Let $\overline{X}_n = (\overline{X}_{1n}, \ldots, \overline{X}_{kn})^T$. Then

$$\boxed{\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)}$$

Summary

- $X_n \xrightarrow{\mathbb{P}} X$: X_n converges to X in probability, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$
- $X_n \xrightarrow{\mathcal{D}} X$: X_n converges to X in distribution, if for all x for which F is continuous
$$\lim_{n\to\infty} F_n(x) = F(x)$$
- $X_n \xrightarrow{\mathbb{P}} X$ implies that $X_n \xrightarrow{\mathcal{D}} X$
- The Law of Large Numbers: Let $X_1, X_2, \ldots$ be an i.i.d. sample and let $\mu = \mathbb{E}[X_1]$. Then
$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu}$$
- The Central Limit Theorem: Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Then
$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

Math 408 - Mathematical Statistics

Lecture 11. Probability Theory: an Overveiw

February 11, 2013

The starting point in developing the probability theory is the notion of a sample space = the set of possible outcomes.

Definition

- The **sample space** Ω is the set of possible outcomes of an "experiment"
- Points $\omega \in \Omega$ are called **realizations**
- **Events** are subsets of Ω

Next, to every event $A \subset \Omega$, we assign a real number $\mathbb{P}(A)$, called the probability of A. We call function $\mathbb{P} : \{\text{subsets of } \Omega\} \to \mathbb{R}$ a probability distribution.

Function $\mathbb{P}$ is not arbitrary, it satisfies several natural properties (called axioms of probability):

1. $0 \leq \mathbb{P}(A) \leq 1$ (Events range from never happening to always happening)
2. $\mathbb{P}(\Omega) = 1$ (Something must happen)
3. $\mathbb{P}(\emptyset) = 0$ (Nothing never happens)
4. $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$ (A must either happen or not-happen)
5. $\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

Statistical Independence

Definition

Two events A and B are **independent** if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

Independence can arise in two distinct ways:

1. We explicitly assume that two events are independent.
2. We derive independence of A and B by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.

Conditional Probability

Definition

If $\mathbb{P}(A) > 0$, then the conditional probability of B given A is

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}}$$

Useful Interpretation:

Think of $\mathbb{P}(B|A)$ as the

fraction of times B occurs among those in which A occurs

Properties of Conditional Probabilities:

1. For any fixed A such that $\mathbb{P}(A) > 0$, $\mathbb{P}(\cdot|A)$ is a probability, i.e. it satisfies the rules of probability.
2. In general $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$
3. If A and B are independent then $\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$
   Thus, another interpretation of independence is that knowing A does not change the probability of B.

Law of Total Probability and Bayes' Theorem

Law of Total Probability

Let $A_1, \ldots, A_n$ be a partition of Ω, i.e.

- $\bigcup_{i=1}^{n} A_i = \Omega$ ($A_1, \ldots, A_n$ are *jointly exhaustive* events)
- $A_i \bigcap A_j = \emptyset$ for $i \neq j$ ($A_1, \ldots, A_n$ are *mutually exclusive* events)
- $\mathbb{P}(A_i) > 0$

Then for any event B

$$\boxed{\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Bayes' Theorem

Conditional probabilities can be inverted. That is,

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}}$$

Random Variables

We need the random variables to link sample spaces and events to data.

Definition

A random variable is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

This mapping induces probability on $\mathbb{R}$ from Ω as follows:
Given a random variable X and a set $A \subset \mathbb{R}$, define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and let

$$\boxed{\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})}$$

Definition

The cumulative distribution function (CDF) $F_X : \mathbb{R} \to [0, 1]$ is defined by

$$\boxed{F_X(x) = \mathbb{P}(X \leq x)}$$

CDF contains all the information about the random variable

Properties of CDFs

Theorem

A function $F : \mathbb{R} \to [0,1]$ is a CDF for some random variable if and only if it satisfies the following three conditions:

1. *F is non-decreasing:*

$$x_1 < x_2 \quad \Rightarrow \quad F(x_1) \leq F(x_2)$$

2. *F is normalized:*

$$\lim_{x \to -\infty} F(x) = 0 \quad \textit{and} \quad \lim_{x \to +\infty} F(x) = 1$$

3. *F is right-continuous:*

$$\lim_{y \to x+0} F(y) = F(x)$$

Discrete Random Variables

Definition

X is **discrete** if it takes countable many values $\{x_1, x_2, \ldots\}$.
We define the **probability mass function** (PMF) for X by

$$\boxed{f_X(x) = \mathbb{P}(X = x)}$$

Relationships between CDF and PMF:

- The CDF of X is related to the PMF f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

- The PMF f_X is related to the CDF F_X by

$$f_X(x) = F_X(x) - F_X(x^-) = F_X(x) - \lim_{y \to x-0} F(y)$$

Continuous Random Variables

Definition

A random variable is **continuous** if there exists a function f_X such that

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$, and
- For every $a \leq b$

$$\boxed{P(a < X \leq b) = \int_a^b f_X(x)dx}$$

- The function $f_X(x)$ is called the probability density function (PDF)
- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)dt} \qquad \boxed{f_X(x) = F_X'(x)}$$

Transformation of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X.
Let $Y = r(X)$ be a function of X.

Q: How to compute the PDFand CDF of Y?

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$
2. Find the CDF $F_Y(y)$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \in A_y) = \int_{A_y} f_X(x)dx$$

3. The PDF is then $f_Y(y) = F'_Y(y)$

Important Fact: When r is strictly monotonic, then r has an inverse $s = r^{-1}$ and

$$\boxed{f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|}$$

Joint Distributions

- Discrete Case

Definition

Given a pair of discrete random variables X and Y, their **joint PMF** is defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$$

- Continuous Case

Definition

A function $f_{X,Y}(x,y)$ is called the **joint PDF** of continuous random variables X and Y if

- $f_{X,Y}(x,y) \geq 0, \qquad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$
- For any set $A \subset \mathbb{R} \times \mathbb{R}$

$$\mathbb{P}((X,Y) \in A) = \int\int_A f_{X,Y}(x,y)dxdy$$

The **joint CDF** of X and Y is defined as $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$

Marginal Distributions

- Discrete Case
  If X and Y have joint PMF $f_{X,Y}$, then the **marginal PMF** of X is

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

  Similarly, the **marginal PMF** of Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f_{X,Y}(x, y)$$

- Continuous Case
  If X and Y have joint PDF $f_{X,Y}$, then the **marginal PDFs** of X and Y are

$$f_X(x) = \int f_{X,Y}(x, y)dy \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x, y)dx$$

Independent Random Variables

Definition

Two random variables X and Y are **independent** if, for every A and B

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Criterion of independence:

Theorem

Let X and Y have joint PDF/PMF $f_{X,Y}$. Then X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Conditional Distributions

- Discrete Case
  The **conditional PMF**:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Continuous Case
  The **conditional PDF** is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Then,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$$

Expectation and its Properties

The expectation (or mean) of a random variable X is the average value of X.

Definition

The **expected value**, or **mean**, or **first moment** of X is

$$\boxed{\mu_X \equiv \mathbb{E}[X] = \begin{cases} \sum_x x f_X(x), & \text{if } X \text{ is discrete} \\ \int x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}}$$

assuming that the sum (or integral) is well-defined.

- Let $Y = r(X)$, then $\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) f_X(x) dx$
- If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

- Let $X_1, \ldots, X_n$ be independent random variables. Then,

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

Variance and its Properties

The variance measures the "spread" of a distribution.

Definition

Let X be a random variance with mean μ_X.
The **variance** of X, denoted $\mathbb{V}[X]$ or σ_X^2, is defined by

$$\boxed{\sigma_X^2 \equiv \mathbb{V}[X] = \mathbb{E}[(X-\mu_X)^2] = \begin{cases} \sum_x (x-\mu_X)^2 f_X(x), & \text{if } X \text{ is discrete} \\ \int (x-\mu_X)^2 f_X(x)dx, & \text{if } X \text{ is continuous} \end{cases}}$$

The **standard deviation** is $\sigma_X = \sqrt{\mathbb{V}[X]}$

Important Properties of $\mathbb{V}[X]$:

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mu_X^2$
- If a and b are constants, then $\mathbb{V}[aX+b] = a^2\mathbb{V}[X]$
- If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i]$$

Covariance and Correlation

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the **covariance** between X and Y by

$$\boxed{\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}$$

and the **correlation** by

$$\boxed{\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}}$$

Properties of Covariance and Correlation

- The covariance satisfies (useful in computations):

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1$$

- If $Y = aX + b$ for some constants a and b, then

$$\rho(X, Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0 \end{cases}$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0$. The converse is not true.
- For random variables $X_1, \ldots, X_n$

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

Conditional Expectation and Conditional Variance

- The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_x x f_{X|Y}(x|y), & \text{discrete case;} \\ \int x f_{X|Y}(x|y)dx, & \text{continuous case.} \end{cases}$$

  - $\mathbb{E}[X]$ is a number
  - $\mathbb{E}[X|Y=y]$ is a function of y
  - $\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y=y]$ when $Y = y$
- The Rule of Iterated Expectations

$$\mathbb{E}\mathbb{E}[Y|X] = \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}\mathbb{E}[X|Y] = \mathbb{E}[X]$$

- The conditional variance of X given $Y = y$ is

$$\mathbb{V}[X|Y=y] = \mathbb{E}[(X - \mathbb{E}[X|Y=y])^2|Y=y]$$

  - $\mathbb{V}[X]$ is a number
  - $\mathbb{V}[X|Y=y]$ is a function of y
  - $\mathbb{V}[X|Y]$ is the random variable whose value is $\mathbb{V}[X|Y=y]$ when $Y = y$
- For random variables X and Y

$$\mathbb{V}[X] = \mathbb{E}\mathbb{V}[X|Y] + \mathbb{V}\mathbb{E}[X|Y]$$

Inequalities

- Markov inequality: If X is a non-negative random variable, then for any $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

- Hoeffding inequality: Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$, then for any $\varepsilon > 0$

$$\mathbb{P}(|\overline{X}_n - p| \geq a) \leq 2e^{-2na^2}$$

- Cauchy-Schwarz inequality: If X and Y have finite variances, then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

- Jensen Inequality:
  - If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
  - If g is concave, then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$

Convergence of Random Variables

There are two main types of convergence: convergence in probability and convergence in distribution.

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

- X_n **converges** to X **in probability**, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

- X_n **converges** to X **in distribution**, written $X_n \xrightarrow{\mathcal{D}} X$, if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

for all x for which F is continuous.

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \quad \text{implies that} \quad X_n \xrightarrow{\mathcal{D}} X}$$

Law of Large Numbers and Central Limit Theorem

The LLN says that the mean of a large sample is close to the mean of the distribution.

The Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu} \quad \text{as } n \to \infty$$

The CLT says that $\overline{X}_n$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Then

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)} \quad \text{as } n \to \infty$$

The Central Limit Theorem

The central limit theorem tells us that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \ \dot{\sim} \ \mathcal{N}(0,1)$$

However, in applications, we rarely know σ. We can estimate σ^2 from $X_1, \ldots, X_n$ by sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Question: If we replace σ with S_n is the central limit theorem still true?

Answer: Yes!

Theorem

Assume the same conditions as in the CLT. Then,

$$\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)} \quad \text{as } n \to \infty$$

Multivariate Central Limit Theorem

Let $X_1, \dots, X_n$ be i.i.d. random vectors with mean μ and covariance matrix Σ:

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_{1i}] \\ \mathbb{E}[X_{2i}] \\ \vdots \\ \mathbb{E}[X_{ki}] \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \mathbb{V}[X_{1i}] & \mathrm{Cov}(X_{1i}, X_{2i}) & \dots & \mathrm{Cov}(X_{1i}, X_{ki}) \\ \mathrm{Cov}(X_{2i}, X_{1i}) & \mathbb{V}[X_{2i}] & \dots & \mathrm{Cov}(X_{2i}, X_{ki}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{ki}, X_{1i}) & \dots & \mathrm{Cov}(X_{ki}, X_{k-1i}) & \mathbb{V}[X_{ki}] \end{pmatrix}$$

Let $\overline{X}_n = (\overline{X}_{1n}, \dots, \overline{X}_{kn})^T$. Then

$$\boxed{\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)} \quad \text{as } n \to \infty$$

Math 408 - Mathematical Statistics

Lecture 12. Introduction to Survey Sampling

February 15, 2013

Agenda

- Goals of Survey Sampling
- Population Parameters
- Simple Random Sampling
- Estimation of the population mean
- Summary

Survey Sampling

Sample surveys are use to obtain information about a large population.
The purpose of **survey sampling** is to reduce the cost and the amount of work that it would take to survey the entire population.

**By a small sample
we may judge of the whole piece**

Miguel de Cervantes
"Don Quixote"

Familiar Examples of Survey Sampling:

- the cook in the kitchen taking a spoonful of soup to determine its taste
- the brewer needing only a sip of beer to test its quality

History of Survey Sampling

The first known attempt to make statements about a population using only information about part of it was made by the English merchant John Graunt. In his famous tract (Graunt, 1662) he describes a method to estimate the population of London based on partial information. John Graunt has frequently been merited as the founder of demography.

Captain John Graunt

The second time a survey-like method was applied was more than a century later. Pierre Simon Laplace realized that it was important to have some indication of the accuracy of the estimate of the French population (Laplace, 1812).

Recommended Reading: "The rise of survey sampling," by J. Bethlehem (2009).

Survey Sampling: Population Parameters

Suppose that the target population is of size N (N is very large) and a numerical value of interest x_i is associated with i^{th} member of the population, $i = 1, \dots, N$.

Examples:

- $x_i =$ age, weight, etc.
- $x_i = 1$ if some characteristic is present, and $x_i = 0$ otherwise.

There are two "standard" parameters of population that we are typically interested:

Definition

- **Population mean**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- **Population variance**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Simple Random Sampling

Important Remark:

Note that μ and σ^2 are not random. They are some fixed unknown parameters. We want to estimate them by picking n out of N members of the population and constructing estimates of μ and σ^2 based only on these n members.

The most elementary form of sampling from a population is **simple random sampling**.

Definition

In Simple Random Sampling, each member is chosen entirely by chance and, therefore, each member has an equal chance of being included in the sample; each particular sample of size n has the same probability of occurrence.

Let $X_1, \ldots, X_n$ be the sample drawn from the population.

Important Remark: Each X_i is a random variable:

- X_i is the value of the i^{th} element of the sample that was randomly chosen from the population
- x_i is the value of the i^{th} member of the population

Estimate

We will consider the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an **estimate** of the population mean μ. Since X_i are random, $\overline{X}_n$ is also random. Distribution of $\overline{X}_n$ is called its sampling distribution.The sampling distribution of $\overline{X}_n$ determines how accurately $\overline{X}_n$ estimates μ: the more tightly the sampling distribution is centered on μ, the better the estimate.

Our goal: is to investigate the sampling distribution of $\overline{X}_n$

Since $\overline{X}_n$ depends on X_i, let us start with examining the distribution of a single sample element X_i.

Basic Lemma

Lemma

Denote the distinct values assumed by the population members by $\xi_1, \ldots, \xi_m$*,* $m \leq N$*, and denote the number of population members that have the value* ξ_i *by* n_i*. Then* X_i *is a discrete random variable with probability mass function*

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N} \tag{1}$$

Also

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2 \tag{2}$$

$\overline{X}_n$ is an unbiased estimator of μ

Theorem

With simple random sampling,

$$\mathbb{E}[\overline{X}_n] = \mu \tag{3}$$

This result can be interpreted as follows: "on average" $\overline{X}_n = \mu$

Definition

Suppose we want to estimate a parameter θ by a function $\hat{\theta}$ of the sample $X_1, \ldots, X_n$,

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

The estimator $\hat{\theta}$ is called **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$

Thus, $\overline{X}_n$ is an unbiased estimator of μ

Summary

- Sample surveys are used to obtain information about a large population
- Population parameters: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We use sample mean $\overline{X}_n$ to estimate the population mean μ.
  - μ is unknown fixed parameter
  - $\overline{X}_n$ is random
- Properties of the sample element X_i:

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N} \qquad \mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2$$

- $\overline{X}_n$ is an unbiased estimator of μ

$$\mathbb{E}[\overline{X}_n] = \mu$$

- Our next goal is to study the sampling distribution of $\overline{X}_n$.

Math 408 - Mathematical Statistics

Lecture 13-14. The Sample Mean and the Sample Variance Under Assumption of Normality

February 20, 2013

Framework

Let $X_1, \ldots, X_n$ be a sample drawn from a population.

Suppose that the population is "Gaussian" $\boxed{X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)}$

We want to estimate population parameters μ and σ^2.

Definition

- The **sample mean** is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- The **sample variance** is $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$

Theorem

$\overline{X}_n$ and S_n^2 are **unbiased estimators** *of μ and σ^2, respectively,*

$$\mathbb{E}[\overline{X}_n] = \mu, \qquad \mathbb{E}[S_n^2] = \sigma^2$$

Our goal: to describe distributions of $\overline{X}_n$ and S_n^2

Distribution of $\overline{X}_n$

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Distribution of S_n^2

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

The χ^2-distribution

Definition

Let $Z_1, \dots, Z_n$ be independent standard normal variables,

$$Z_1, \dots, Z_n \sim N(0,1)$$

Then the distribution of

$$Q = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

is called the χ^2**-distribution** with n **degrees of freedom**,

$$Q \sim \chi_n^2$$

- Probability Density Function:

$$\pi(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

  - $x \geq 0$
  - Γ is the gamma function $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$

The χ^2-distribution

The χ^2-distribution is especially important in hypothesis testing.

Nice Properties:

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

$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

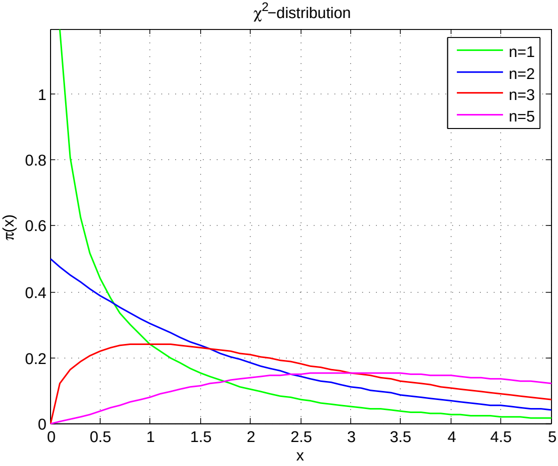
  and

$$\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$$

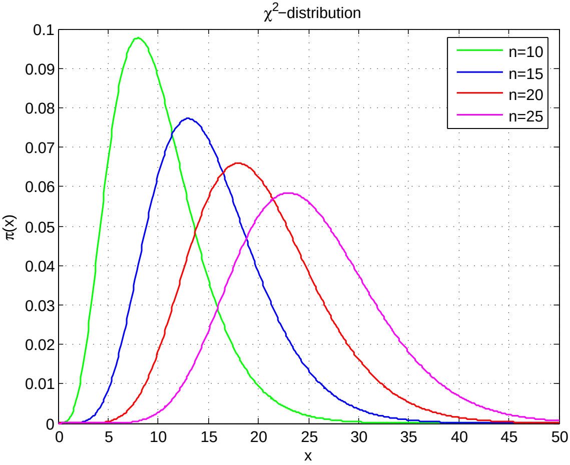
- If $U \sim \chi_n^2$ and $V \sim \chi_m^2$, and U and V are independent, then

$$U+V \sim \chi_{n+m}^2$$

Graph of the χ^2_n PDF: small n



Graph of the χ_n^2 PDF: large n



- CLT: χ_n^2 converges to a normal distribution as $n \to \infty$
- $\chi_n^2 \to \mathcal{N}(n, 2n)$, as $n \to \infty$
- When $n > 50$, for many practical purposes, $\chi_n^2 = \mathcal{N}(n, 2n)$

Distribution of S_n^2

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Proof: is based on moment-generating functions...

Moment-generating functions

Definition

The moment-generating function (MGF) of a random variable $X \sim f(x)$ is

$$M(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

(if the expectation is defined)

Important Properties of MGFs:

- If $\exists\, \varepsilon > 0$ such that $M(t)$ exists for all $t \in (-\varepsilon, \varepsilon)$, then $M(t)$ uniquely determines the probability distribution, $M(t) \rightsquigarrow f(x)$.
- If $M(t)$ exists in an open interval containing zero, then

$$M^{(r)}(0) = \mathbb{E}[X^r] \qquad \text{(hence the name)}$$

To find moments $\mathbb{E}[X^r]$, we must do integration.
Knowing the MGF allows to replace integration by differentiation.

Moment-generating functions

Important Properties of MGFs: (continuation)

- If X has the MGF $M_X(t)$ and $Y = a + bX$, then

$$M_Y(t) = e^{at} M_X(bt)$$

- If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

- If X and Y have a joint distribution, then their joint MGF is defined as

$$M_{X,Y}(s,t) = \mathbb{E}[e^{sX+tY}]$$

  X and Y are independent if and only if

$$M_{X,Y}(s,t) = M_X(s) M_Y(t)$$

Moment-generating functions: Limitations and Examples

The major limitation of the moment-generating function is that it may not exist.

In this case, the characteristic function may be used:

$$\phi(t) = \mathbb{E}[e^{itX}]$$

<u>Examples:</u>

- $\mathcal{N}(\mu, \sigma^2)$:

$$M(t) = e^{\mu t} e^{\sigma^2 t^2/2}$$

- χ_n^2:

$$M(t) = (1 - 2t)^{-n/2}$$

Distribution of S_n^2

Theorem

If $X_1, \dots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Bringing the t-distribution into the Game

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

The t-distribution

Definition

Let $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2_n$, and Z and U are independent. Then the distribution of

$$T = \frac{Z}{\sqrt{U/n}}$$

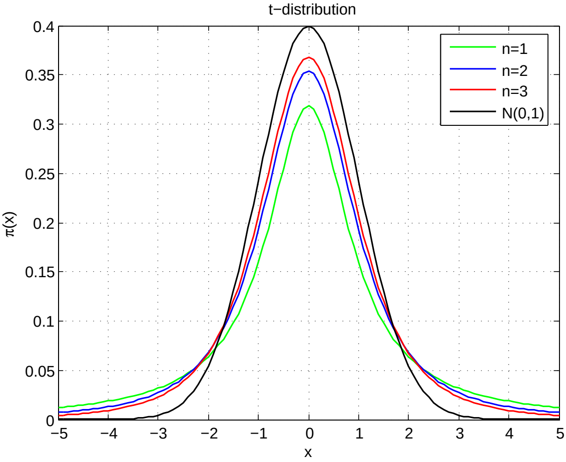
is called the t-**distribution** with n **degrees of freedom**.

- Probability Density Function:

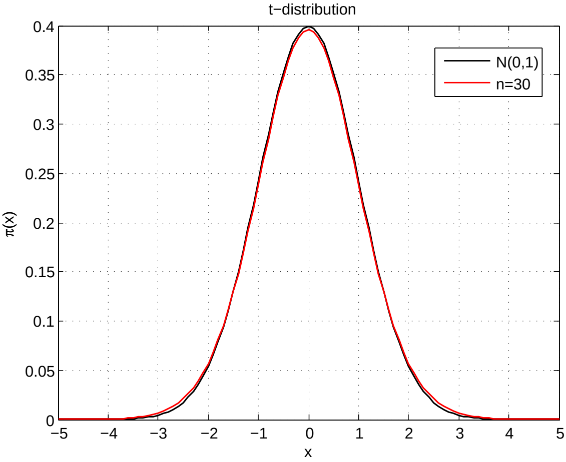
$$\pi(x) = \frac{\Gamma\left((n+1)/2\right)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

- The t-distribution is symmetric about zero, $\pi(x) = \pi(-x)$
- As $n \to \infty$, the t-distribution tends to the standard normal distribution. In fact, when $n > 30$, the two distributions are very close.

Graph of the t-distribution PDF: small n



Graph of the t-distribution PDF: large n



Bringing the t-distribution into the Game

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

Summary

Under Assumption of Normality, $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$,

$$\text{the sample mean:} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{the sample variance:} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

have the following properties:

- $\boxed{\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)}$

- $\boxed{\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}}$ $\qquad \chi^2_n = \mathcal{N}(0,1)^2 + \ldots + \mathcal{N}(0,1)^2$

- $\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}}$ $\qquad t_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/n}}$

Math 408 - Mathematical Statistics

Lecture 15. Accuracy of estimation of the population mean $\overline{X}_n \approx \mu$

February 25, 2013

In Lecture 12, we discussed the basic mathematical framework of survey sampling:

- We have the target population of size N (N is very large).
- A numerical value of interest x_i (age, weight, income, etc) is associated with i^{th} member of the population.
- We are interested in population parameters:
  - Population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We estimate μ by the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ is a sample drawn from the population using the simple random sampling.

We proved that $\overline{X}_n$ is an unbiased estimate of μ:

$$\boxed{\mathbb{E}[\overline{X}_n] = \mu}$$

In other words, on average $\overline{X}_n \approx \mu$.

Our next goal is to investigate how variable $\overline{X}_n$ is

As a measure of the dispersion of $\overline{X}_n$ about μ, we will use the standard deviation of $\overline{X}_n$, $\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$.

Thus, we want to find

$$\boxed{\mathbb{V}[\overline{X}_n] = ?}$$

$$\mathbb{V}[\overline{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right]$$

Remark: If sampling were done with replacement then X_i would be independent, and we would have:

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

In simple random sampling, we do sampling without replacement.
This induces dependence among X_i. And therefore

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \neq \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i]$$

Recall Lecture 6:

$$\mathbb{V}\left[\sum_{i=1}^{n} \alpha_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \mathrm{Cov}(X_i, X_j)$$

Thus, we have:

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

So, we need to find $\mathrm{Cov}(X_i, X_j)$.

Lemma

If $i \neq j$, then the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Theorem

The variance of $\overline{X}_n$ is given by

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

Important observations:

- If $n << N$, then

$$\mathbb{V}[\overline{X}_n] \approx \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

  $\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction.

- To double the accuracy of $\mu \approx \overline{X}_n$, the sample size must be quadrupled
- If σ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if σ is large, then a larger sample will be required to obtain the same accuracy.

Summary

- The main result of this lecture is the expression for the variance of $\overline{X}_n$:

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

- The corresponding standard deviation

$$\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$$

measures the dispersion of $\overline{X}_n$ about μ.

Math 408 - Mathematical Statistics

Lecture 16. Estimation of the Population Variance σ

February 27, 2013

Agenda

- Why do we need to estimate σ?
- How can we estimate σ?
- Summary

The Need of Estimation of σ

We know that the sample mean $\overline{X}_n$ is an unbiased estimate of the population mean μ:

$$\mathbb{E}[\overline{X}_n] = \mu$$

Moreover, the accuracy of the approximation $\overline{X}_n \approx \mu$ can be measured by the standard deviation of $\overline{X}_n$ (also called "standard error"):

$$\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}, \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}, \quad \text{if } n \ll N \tag{1}$$

where σ is the population variance

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

<u>Q:</u> What is the main drawback of (1)?

<u>A:</u> We can't use (1) since σ is unknown.
To use (1), σ must be estimated from the sample $X_1, \ldots, X_n$.

Estimation of σ

It seems natural to use the following estimate

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

However, this estimate is biased.

Theorem

The expected value of $\hat{\sigma}_n^2$ is given by

$$\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn - N}{Nn - n}$$

Important Remark:

- Since $\frac{Nn-N}{Nn-n} < 1$, we have $\mathbb{E}[\hat{\sigma}_n^2] < \sigma^2$
  Therefore, $\hat{\sigma}_n^2$ tends to underestimate σ^2

Corollaries

Corollary

Since $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn-N}{Nn-n}$,

$$\hat{\sigma}^2_{n,\text{unbiased}} = \frac{Nn-n}{Nn-N}\hat{\sigma}_n^2$$

is an unbiased estimate of σ^2

Recall that

$$\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

In practice, σ is unknown, so we need to estimate it.

Corollary

An unbiased estimate of $\mathbb{V}[\overline{X}_n]$ *is*

$$\boxed{s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n}\frac{Nn-n}{Nn-N}\left(1 - \frac{n-1}{N-1}\right)}$$

Summary

Let us summarize what we have learned about estimation of population parameters:

- **Population mean** μ
  - Unbiased estimate:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - Variance of estimate

$$\mathbb{V}[\overline{X}_n] \equiv \sigma^2_{\overline{X}_n} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

  - Estimated variance

$$\sigma^2_{\overline{X}_n} \approx s^2_{\overline{X}_n} = \frac{\hat{\sigma}^2_n}{n}\frac{Nn-n}{Nn-N}\left(1 - \frac{n-1}{N-1}\right)$$

- **Population variance** σ
  - Unbiased estimate:

$$\hat{\sigma}^2_{n,\mathrm{unbiased}} = \frac{Nn-n}{Nn-N}\hat{\sigma}^2_n, \qquad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Conclusion

In simple random sampling, we can not only form estimate of unknown population parameter (e.g. μ), but also obtain the likely size of errors of these estimates.In other words, we can obtain the estimate of a parameter as well as the estimate of the error of that estimate

Math 408 - Mathematical Statistics

Lecture 17. The Normal Approximation to the Distribution of $\overline{X}_n$

March 1, 2013

Agenda

- Normal Approximation (theoretical result)
- Approximation of the Error Probabilities (application 1)
- Confidence Intervals (application 2)
- Example: Hospitals
- Summary

We previous Lectures, we found the mean and the variance of the sample mean:

$$\mathbb{E}[\overline{X}_n] = \mu \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Ideally, we would like to know the **entire distribution** of $\overline{X}_n$ (sampling distribution) since it would tell us everything about the random variable $\overline{X}_n$

<u>Reminder:</u>

If $X_1, \ldots, X_n$ are i.i.d. with the common mean μ and variance σ^2, then the sample mean $\overline{X}_n$ has the following properties:

1. $\mathbb{E}[\overline{X}_n] = \mu, \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$
2. CLT:
$$\mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z), \qquad \text{as } n \to \infty$$
where $\Phi(z)$ is the CDF of $\mathcal{N}(0,1)$

<u>Q:</u> Can we use these results to obtain the distribution of $\overline{X}_n$?
<u>A:</u> No. In simple random sampling, X_i are not independent.
Moreover, it makes no sense to have n tend to infinity while N is fixed.

Nevertheless, it can be shown that if n is large, but still small relative to N, then $\overline{X}_n$ is **approximately normally distributed**

$$\boxed{\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})} \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

How can we use this results?

Suppose we want to find the probability that the error made in estimating μ by $\overline{X}_n$ is less than $\varepsilon > 0$. In symbols, we want to find

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) = ?$$

Theorem

From $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})$ it follows that

$$\boxed{\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1}$$

Confidence Intervals

Let $\alpha \in [0, 1]$

Definition

A $100(1-\alpha)\%$ **confidence interval** for a population parameter θ is a random interval calculated from the sample, which contains θ with probability $1-\alpha$.

Interpretation:

If we were to take many random samples and construct a confidence interval from each sample, then about $100(1-\alpha)\%$ of these intervals would contain θ.

Our goal: to construct a confidence interval for μ

Let z_α be that number such that the area under the standard normal density function to the right of z_α is α. In symbols, z_α is such that

$$\Phi(z_\alpha) = 1 - \alpha$$

Useful property:

$$z_{1-\alpha} = -z_\alpha$$

Confidence interval for μ

Theorem

An (approximate) $100(1-\alpha)\%$ *confidence interval for* μ *is*

$$(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n})$$

That is the probability that μ *lies in that interval is approximately* $1-\alpha$

$$\boxed{\mathbb{P}(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}) \approx 1-\alpha}$$

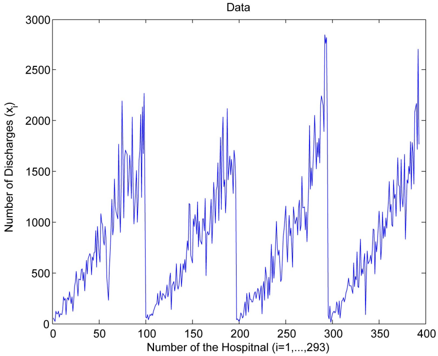
Remarks:

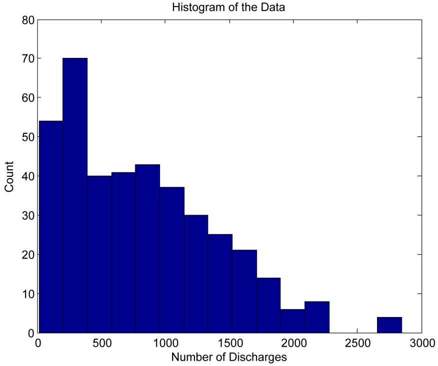
- This confidence interval is random. The probability that it covers μ is $(1-\alpha)$
- In practice, $\alpha = 0.1, 0.05, 0.01$ (depends on a particular application)
- Since $\sigma_{\overline{X}_n}$ is not known (it depends on σ), $s_{\overline{X}_n}$ is used instead of $\sigma_{\overline{X}_n}$

Example: Hospitals

Data: Herkson (1976):

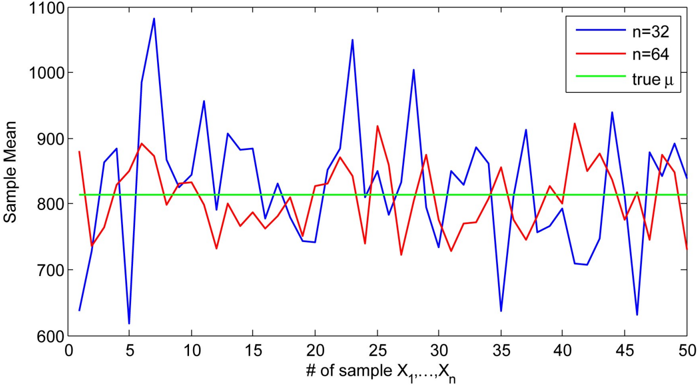
- The population consists of $N = 393$ short-stay hospitals
- Let x_i be the number of patients discharged from the i^{th} hospital during January 1968.





Example: Hospitals

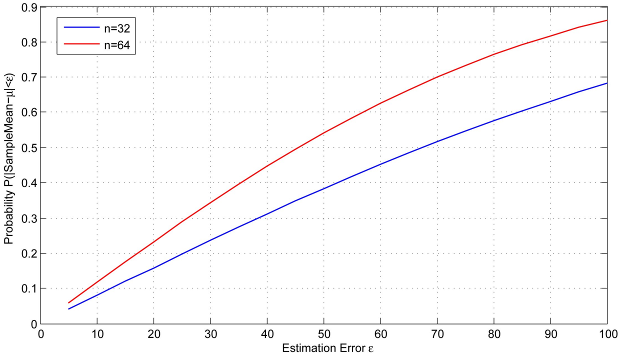
- Population mean $\mu = 814.6$, and population variance $\sigma^2 = (589.7)^2$
- Let us consider two case $n_1 = 32$ and $n_2 = 64$.

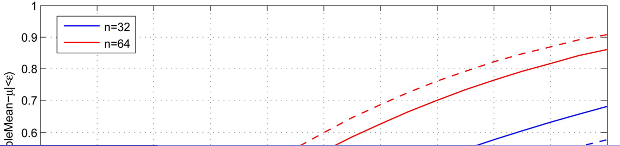


- True std of $\overline{X}_n$: $\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$, $\sigma_{\overline{X}_{32}} = 100$, $\sigma_{\overline{X}_{64}} = 67.5$

Example: Hospitals

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$



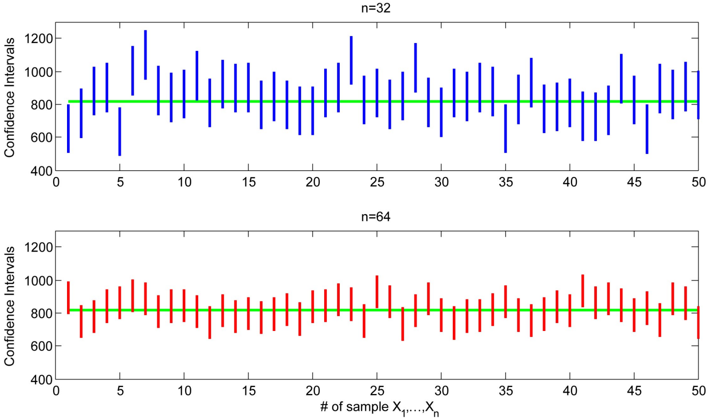


Example: Hospitals

$100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}\right)$$

$\alpha = 0.1$:



Interval width: 329.1 for $n = 32$ and 222.2 for $n = 64$

Summary

- The sample mean is approximately normal

$$\boxed{\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})} \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

- Probability of error

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

- $100(1-\alpha)\%$ confidence interval for μ is

$$(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n})$$

Math 408 - Mathematical Statistics

Lecture 18. Estimation of a Ratio and the δ-method

March 4, 2013

Ratio and its Estimate

Suppose that for each member of a population, two values are measured:

$$i^{\text{th}} \text{ member} \rightsquigarrow (x_i, y_i)$$

We are interested in the following **ratio**:

$$\boxed{r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}}$$

Ratios arise frequently in sample surveys.

Example:

Households are sampled. If y_i is the number of unemployed males in the i^{th} household, and x_i is the total number of males in the i^{th} household, then r is the proportion of unemployed males.

Estimate of a Ratio

Let $\begin{pmatrix} X_1 & \dots & X_n \\ Y_1 & \dots & Y_n \end{pmatrix}$ be a sample from a population.
Then the natural estimate of

$$r = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \frac{\frac{1}{N}\sum_{i=1}^N y_i}{\frac{1}{N}\sum_{i=1}^N x_i} = \frac{\mu_y}{\mu_x}$$

is

$$\boxed{R_n = \frac{\overline{Y}_n}{\overline{X}_n}}$$

Our goal: to derive expressions for $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$

Technical problem: since R_n a nonlinear function of $\overline{X}_n$ and $\overline{Y}_n$, we can't find $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$ in closed form.

Idea: To approximate $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$ using the δ-**method**.

The δ-method

In many applications, the following scenario is typical:

Problem

X is a random variable, μ_x and σ_X^2 are known. The problem is to find the mean and variance of $Y = f(X)$, where f is some (typically nonlinear) function.

The δ-**method** utilizes a strategy that is often used in applied mathematics: when confronted with a nonlinear problem that we can't solve, we linearize.

In the δ-method, the linearization is carried out through a Taylor series expansion of f about μ_X:

$$Y = f(X) \approx f(\mu_X) + (X - \mu_X)f'(\mu_X)$$

We thus obtain the first order approximations:

$$\boxed{\mu_Y \approx f(\mu_X)} \qquad \boxed{\sigma_Y^2 \approx (f'(\mu_X))^2\sigma_X^2}$$

The δ-method

To obtain a better approximation for μ_Y, we can use the Taylor series expansion to the 2^{nd} order:

$$Y = f(X) \approx f(\mu_X) + (X - \mu_X)f'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 f''(\mu_X)$$

Then the second order approximations for μ_Y is

$$\boxed{\mu_Y \approx f(\mu_X) + \frac{1}{2}\sigma_X^2 f''(\mu_X)}$$

We can similarly proceed in the case of two random variables X and Y:

Problem

Suppose that $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY} = Cov(X, Y)$ are known.
The problem is to find μ_Z and σ_Z^2, where $Z = f(X, Y)$.

The δ-method

Using the Taylor series expansion to the first order:

$$Z = f(X,Y) \approx f(\mu) + (X-\mu_X)\frac{\partial f}{\partial x}(\mu) + (Y-\mu_Y)\frac{\partial f}{\partial y}(\mu), \qquad \mu = (\mu_X, \mu_Y)$$

Therefore,

$$\boxed{\mu_Z \approx f(\mu)} \qquad \boxed{\sigma_Z^2 \approx \sigma_X^2\left(\frac{\partial f}{\partial x}(\mu)\right)^2 + \sigma_Y^2\left(\frac{\partial f}{\partial y}(\mu)\right)^2 + 2\sigma_{XY}\frac{\partial f}{\partial x}(\mu)\frac{\partial f}{\partial y}(\mu)}$$

To obtain a better approximation for μ_Z, we can use the Taylor series expansion to the second order.

$$\boxed{\mu_Z \approx f(\mu) + \frac{1}{2}\sigma_X^2\frac{\partial^2 f}{\partial x^2}(\mu) + \frac{1}{2}\sigma_Y^2\frac{\partial^2 f}{\partial y^2}(\mu) + \sigma_{XY}\frac{\partial^2 f}{\partial x \partial y}(\mu)}$$

The δ-method: special case $Z = Y/X$

Example

If $Z = Y/X$, then

$$\boxed{\mu_Z \approx \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X^2}\left(\sigma_X^2 \frac{\mu_Y}{\mu_X} - \sigma_{XY}\right)} \tag{1}$$

$$\boxed{\sigma_Z^2 \approx \frac{1}{\mu_X^2}\left(\sigma_X^2 \frac{\mu_Y^2}{\mu_X^2} + \sigma_Y^2 - 2\sigma_{XY}\frac{\mu_Y}{\mu_X}\right)} \tag{2}$$

Approximations of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$

The estimate of $r = \frac{\mu_y}{\mu_x}$ is

$$R_n = \frac{\overline{Y}_n}{\overline{X}_n}$$

To use the δ-method to approximate $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$, we need to know $\mu_{\overline{X}_n}, \mu_{\overline{Y}_n}, \sigma^2_{\overline{X}_n}, \sigma^2_{\overline{Y}_n}$, and $Cov(\overline{X}_n, \overline{Y}_n)$. In previous Lectures, we found that

- $\mu_{\overline{X}_n} = \mu_x$
- $\mu_{\overline{Y}_n} = \mu_y$
- $\sigma^2_{\overline{X}_n} = \frac{\sigma_x^2}{n}\left(1 - \frac{n-1}{N-1}\right)$
- $\sigma^2_{\overline{Y}_n} = \frac{\sigma_y^2}{n}\left(1 - \frac{n-1}{N-1}\right)$

It can be shown that

- $Cov(\overline{X}_n, \overline{Y}_n) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right)$, where σ_{xy} is the population covariance of x and y, $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$.

Approximations of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$

Using approximations (1) and (2) from the δ-method, we obtain

Theorem

The expectation and variance of R_n are given by

$$\boxed{\mathbb{E}[R_n] \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \sigma_{xy})} \tag{3}$$

$$\boxed{\mathbb{V}[R_n] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})} \tag{4}$$

In applications, population parameters $\mu_x, \sigma_x, \sigma_y, \sigma_{xy}$ are unknown. To compute the **estimated** values of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$, we use (3) and (4) together with

- $r \approx R_n \qquad \mu_x \approx \overline{X}_n$
- $\sigma_x^2 \approx \hat{\sigma}^2_{x,\mathrm{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)^2$
- $\sigma_y^2 \approx \hat{\sigma}^2_{y,\mathrm{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$
- $\sigma_{xy} \approx \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$

Summary

- Ratios $r = \mu_y/\mu_x$ arise frequently in sample surveys
- The natural estimate of r is $R_n = \overline{Y}_n/\overline{X}_n$
- We can find expressions for $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$ using the δ-method:

$$\boxed{\mathbb{E}[R_n] \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \sigma_{xy})}$$

$$\boxed{\mathbb{V}[R_n] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})}$$

- To compute the estimated values of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$, we use:
  - $r \approx R_n \qquad \mu_x \approx \overline{X}_n$
  - $\sigma_x^2 \approx \hat{\sigma}^2_{x,\text{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)^2$
  - $\sigma_y^2 \approx \hat{\sigma}^2_{y,\text{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$
  - $\sigma_{xy} \approx \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$

Math 408 - Mathematical Statistics

Lecture 19. Stratified Random Sampling

March 6, 2013

Agenda

- Definition of the Stratified Random Sampling (StrRS)
- Basic statistical properties of estimate of μ obtained under StrRS
- Neyman Allocation Scheme
- Summary

Stratified Random Sampling

In stratified random sampling (StrRS), the population is partitioned into subpopulations, or strata, which are then independently sampled.

In many applications, stratification is natural.

Example:

In samples of human populations, geographical areas form natural strata.

Reasons for using StrRS:

- We are often interested in obtaining information about each natural subpopulation in addition to information about the whole population.
- Estimates obtained from StrRS can be considerably more accurate than estimates from simple random sampling if
  - population members within each stratum are relatively homogeneous, and
  - there is considerable variation between strata.

Mathematical Framework of StrRS

Suppose there are L strata. Let N_k be the number of population elements in the k^{th} stratum. The total population size is

$$N = \sum_{i=1}^{L} N_k$$

Denote the mean and variance of the k^{th} stratum by μ_k and σ_k^2, respectively. Let $x_i^{(k)}$ denote the i^{th} value in the k^{th} stratum, then the overall population mean

$$\mu = \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k} x_i^{(k)} = \frac{1}{N}\sum_{k=1}^{L} N_k\mu_k = \sum_{k=1}^{L}\frac{N_k}{N}\mu_k = \sum_{k=1}^{L}\omega_k\mu_k, \qquad \omega_k = \frac{N_k}{N}$$

Thus, the overall population mean is

$$\mu = \sum_{k=1}^{L}\omega_k\mu_k, \qquad \omega_k = \frac{N_k}{N},$$

where ω_k is the fraction of the population in the k^{th} stratum.

Mathematical Framework of StrRS

Within each stratum, a simple random sample $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ of size n_k is taken. The sample mean is

$$\overline{X}_{n_k}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i^{(k)}, \qquad k = 1, \ldots, L$$

Since $\mu = \sum_{k=1}^{L} \omega_k \mu_k$, the natural estimate of μ is

$$\boxed{\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)}}$$

Remark:

We use star to distinguish $\overline{X}_n^*$ (obtained from stratified random sampling) from $\overline{X}_n$ (obtained from simple random sampling)

Our goal: to study statistical properties of $\overline{X}_n^*$

In particular, we want to find $\mathbb{E}[\overline{X}_n^*]$ and $\mathbb{V}[\overline{X}_n^*]$

Expectation $\mathbb{E}[\overline{X}_n^*]$

Theorem

$\overline{X}_n^*$ *is an unbiased estimate of* μ,

$$\boxed{\mathbb{E}[\overline{X}_n^*] = \mu}$$

Variance $\mathbb{V}[\overline{X}_n^*]$

Theorem

Under stratified random sampling,

$$\boxed{\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}\left(1 - \frac{n_k - 1}{N_k - 1}\right)}$$

Corollary

If the sampling fractions within each stratum are small, i.e. $n_k/N_k \ll 1$, *then*

$$\boxed{\mathbb{V}[\overline{X}_n^*] \approx \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}}$$

Our next goal: to decide how to choose sample sizes $n_1, \ldots, n_L$ efficiently

Neyman Allocation Scheme

So, it was shown that (neglecting the sampling fractions $n_k/N_k \ll 1$)

$$\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}$$

Question:

Suppose that the resources of a survey allow only a total of n units to be sampled. How to choose $n_1, \ldots, n_L$ to minimize $\mathbb{V}[\overline{X}_n^*]$ subject to constraint $\sum n_k = n$?

Optimization problem:

$$\mathbb{V}[\overline{X}_n^*] \to \min \qquad \text{s.t.} \sum_{k=1}^{L} n_k = n \tag{1}$$

Theorem

The sample sizes $n_1, \ldots, n_L$ that solve the optimization problem (1) are given by

$$\boxed{n_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j}} \qquad k = 1, \ldots, L$$

- This optimal allocation scheme is called Neyman allocation

Summary

- Stratified Random Sampling:
  population is partitioned onto strata which are then sampled independently.
- Under stratified random sampling, the estimate of μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)}$$

- The expectation and variance (assuming $n_k/N_k \ll 1$):

$$\boxed{\mathbb{E}[\overline{X}_n^*] = \mu} \qquad \boxed{\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}}$$

- Neyman Allocation Scheme minimizes $\mathbb{V}[\overline{X}_n^*]$ subject to $\sum_{k=1}^{N} n_k = n$:

$$\boxed{n_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j}} \qquad k = 1, \ldots, L$$

Math 408 - Mathematical Statistics

Lecture 20-21.
Neyman Allocation vs Proportional Allocation
and
Stratified Random Sampling vs Simple Random Sampling

March 8-13, 2013

Agenda

- Neyman Allocation and its properties
- Variance of the optimal stratified estimate $\overline{X}^*_{n,opt}$
- Drawbacks of Neyman Allocation
- Proportional Allocation
- Neyman vs Proportional
- Stratified vs Simple
- Summary

Neyman allocation

In Lecture 19, we described the optimal allocation scheme for stratified random sampling, called Neyman allocation. Neyman allocation scheme minimizes variance $\mathbb{V}[\overline{X}_n^*]$ subject to $\sum_{k=1}^N n_k = n$.

Theorem

The sample sizes $n_1, \ldots, n_L$ that solve the optimization problem

$$\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k} \to \min \qquad \text{s.t.} \sum_{k=1}^{L} n_k = n$$

are given by

$$\boxed{\hat{n}_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^L \omega_j \sigma_j}} \qquad k = 1, \ldots, L \tag{1}$$

The theorem says that if $\omega_k \sigma_k$ is large, then the corresponding stratum should be sampled heavily. This is very natural since

- if ω_k is large, then the stratum contains a large portion of the population
- if σ_k is large, then the population values in the stratum are quite variable and, therefore, to estimate μ_k accurately a relatively large sample size must be used

Variance of the optimal stratified estimate

In stratified random sampling, an (unbiased) estimate of μ is

$$\overline{X}^*_n = \sum_{k=1}^{L} \omega_k \overline{X}^{(k)}_{n_k}$$

If Neyman (i.e. optimal) allocation is used ($n_k = \hat{n}_k$), then the optimal stratified estimate of μ, denoted by $\overline{X}^*_{n,opt}$, is

$$\overline{X}^*_{n,opt} = \sum_{k=1}^{L} \omega_k \overline{X}^{(k)}_{\hat{n}_k}$$

Theorem

The variance of the optimal stratified estimate is

$$\mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n}\left(\sum_{k=1}^{L} \omega_k \sigma_k\right)^2$$

Proportional Allocation

There are two main disadvantages of Neyman allocation:

1. Optimal allocations $\hat{n}_k$ depends on σ_k which generally will not be known
2. If a survey measures several values for each population member, then it is usually impossible to find an allocation that is simultaneously optimal for all values

A simple and popular alternative method of allocation is proportional allocation: to choose $n_1, \ldots, n_L$ such that

$$\boxed{\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}}$$

This holds if

$$\boxed{\tilde{n}_k = n\frac{N_k}{N} = n\omega_k} \qquad k = 1, \ldots, L \tag{2}$$

Proportional Allocation

If proportional allocation is used ($n_k = \tilde{n}_k = n\omega_k$), then the corresponding stratified estimate of μ, denoted by $\overline{X}^*_{n,p}$, is

$$\overline{X}^*_{n,p} = \sum_{k=1}^{L} \omega_k \overline{X}^{(k)}_{\tilde{n}_k} = \sum_{k=1}^{L} \omega_k \frac{1}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k} X_i^{(k)} = \frac{1}{n} \sum_{k=1}^{L} \sum_{i=1}^{\tilde{n}_k} X_i^{(k)}$$

Thus, $\overline{X}^*_{n,p}$ is simply the unweighted mean of the sample values.

Theorem

The variance of $\overline{X}^*_{n,p}$ is given by

$$\mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n} \sum_{k=1}^{L} \omega_k \sigma_k^2$$

Neyman vs Proportional

By definition, Neyman allocation is always better than proportional allocation (since Neyman allocation is optimal).

Question: When is it substantially better?

Proposition

$$\mathbb{V}[\overline{X}^*_{n,p}] - \mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k(\sigma_k - \bar{\sigma})^2, \qquad \bar{\sigma} = \sum_{k=1}^{L}\omega_k\sigma_k$$

Therefore,

- if the variances σ_k of the strata are all the same, then proportional allocation is as efficient as Neyman allocation, $\mathbb{V}[\overline{X}^*_{n,p}] = \mathbb{V}[\overline{X}^*_{n,opt}]$
- the more variable σ_k, the more efficient the Neyman allocation scheme

Stratified vs Simple

Let us now compare simple random sampling and stratified random sampling with proportional allocation.

Question: What is more efficient? (more efficient = has smaller variance)

Proposition

$$\mathbb{V}[\overline{X}_n] - \mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n}\sum_{k=1}^{L} \omega_k(\mu_k - \mu)^2$$

Thus, stratified random sampling with proportional allocation always gives a smaller variance than simple random sampling does (providing that the finite population correction is ignored, $(n-1)/(N-1) \approx 0$).

Summary

- The variance of the optimal stratified estimate (Neyman allocation) of μ is

$$\mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n}\left(\sum_{k=1}^{L}\omega_k\sigma_k\right)^2$$

- Neyman allocation is difficult to implement in practice
- Proportional allocation: $\tilde{n}_k = n\frac{N_k}{N} = n\omega_k$
- The variance of the stratified estimate under proportional allocation:

$$\mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k\sigma_k^2$$

- By definition, Neyman allocation is better than proportional allocation, but if the variances σ_k of the strata are all the same, then proportional allocation is as efficient as Neyman allocation
- Stratified random sampling with proportional allocation is always more efficient than simple random sampling.

Math 408 - Mathematical Statistics

Lecture 22. Survey Sampling: an Overview

March 25, 2013

Survey Sampling: What and Why

In **surveys sampling** we try to obtain information about a large population based on a relatively small sample of that population.

The main goal of **survey sampling** is to reduce the cost and the amount of work that it would take to explore the entire population.

First examples: Graunt (1662) and Laplace (1812) used survey sampling to estimate the population of London and France, respectively.

Mathematical Framework

Suppose that the target population is of size N (N is large) and a numerical value of interest x_i (age, weight, income, etc) is associated with i^{th} member of the population, $i = 1, \ldots, N$. Population parameters (quantities we are interested in):

- Population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

There are several ways to sample from a population. We discussed two:

1. Simple Random Sampling

Definition

In Simple Random Sampling, each member is chosen entirely by chance and, therefore, each member has an equal chance of being included in the sample; each particular sample of size n has the same probability of occurrence.

If $X_1, \ldots, X_n$ is the sample drawn from the population, then the sample mean is a natural estimate of the population mean μ:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu$$

2. Stratified Random Sampling

Definition

In Stratified Random Sampling, the population is partitioned into subpopulations, or strata, which are then independently sampled using simple random sampling.

If $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ is the sample drawn from the k^{th} stratum, then the natural estimate of μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)} \approx \mu$$

Statistical Properties of $\overline{X}_n$

Since $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, statistical properties of $\overline{X}_n$ are completely determined by statistical properties of X_i.

Lemma

Denote the distinct values assumed by the population members by $\xi_1, \ldots, \xi_m$, $m \leq N$, and denote the number of population members that have the value ξ_i by n_i. Then X_i is a discrete random variable with probability mass function

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N}$$

Also

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2$$

From this lemma, it follows immediately that $\overline{X}_n$ is an unbiased estimate of μ:

$$\mathbb{E}[\overline{X}_n] = \mu$$

Thus, on average $\overline{X}_n = \mu$.

Statistical Properties of $\overline{X}_n$

The next important question is how variable $\overline{X}_n$ is.

As a measure of the dispersion of $\overline{X}_n$ about μ, we use the standard deviation of $\overline{X}_n$, denoted as $\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$.

Theorem

The variance of $\overline{X}_n$ is given by

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

Important observations:

- If $n << N$, then

$$\mathbb{V}[\overline{X}_n] \approx \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

  $\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction. This factor arises because of dependence among X_i.

Statistical Properties of $\overline{X}_n$

$$\sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}} \tag{1}$$

- To double the accuracy, the sample size must be quadrupled.
- If σ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if σ is large, then a larger sample will be required to obtain the same accuracy.
- We can't use (1) in practice, since σ is unknown. To use (1), σ must be estimated from sample $X_1, \ldots, X_n$.

At first glance, it seems natural to use the following estimate

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \approx \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

However, this estimate is biased.

Statistical Properties of $\overline{X}_n$

Theorem

The expected value of $\hat{\sigma}_n^2$ is given by

$$\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn - N}{Nn - n}$$

In particular, $\hat{\sigma}_n^2$ tends to underestimate σ^2.

Corollary

- *An unbiased estimate of σ^2 is*

$$\hat{\sigma}^2_{n,\mathrm{unbiased}} = \frac{Nn - n}{Nn - N} \hat{\sigma}_n^2$$

- *An unbiased estimate of $\mathbb{V}[\overline{X}_n]$ is*

$$s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n} \frac{Nn - n}{Nn - N} \left(1 - \frac{n-1}{N-1}\right)$$

Normal Approximation to the Distribution of $\overline{X}_n$

So, we know that the sample mean $\overline{X}_n$ is an unbiased estimate of μ, and we know how to approximately find its standard deviation $\sigma_{\overline{X}_n} \approx s_{\overline{X}_n}$.

Ideally, we would like to know the **entire distribution** of $\overline{X}_n$ (sampling distribution) since it would tell us everything about the accuracy of the estimation $\overline{X}_n \approx \mu$

It can be shown that if n is large, but still small relative to N, then $\overline{X}_n$ is **approximately normally distributed**

$$\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n}) \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

From this result, it is easy to find the probability that the error made in estimating μ by $\overline{X}_n$ is less than $\varepsilon > 0$:

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

Confidence Intervals

Let $\alpha \in [0, 1]$

Definition

A $100(1-\alpha)\%$ **confidence interval** for a population parameter θ is a random interval calculated from the sample, which contains θ with probability $1-\alpha$.

Interpretation:

If we were to take many random samples and construct a confidence interval from each sample, then about $100(1-\alpha)\%$ of these intervals would contain θ.

Theorem

An (approximate) $100(1-\alpha)\%$ *confidence interval for* μ *is*

$$(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n})$$

That is the probability that μ *lies in that interval is approximately* $1-\alpha$

$$\boxed{\mathbb{P}(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}) \approx 1-\alpha}$$

Estimation of a Ratio

Suppose that for each member of a population, two values are measured:

$$i^{\text{th}} \text{ member} \rightsquigarrow (x_i, y_i)$$

We are interested in the following **ratio**:

$$r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

Let $\begin{pmatrix} X_1 & \dots & X_n \\ Y_1 & \dots & Y_n \end{pmatrix}$ be a simple random sample from a population.
Then the natural estimate of r is

$$R_n = \frac{\overline{Y}_n}{\overline{X}_n}$$

To obtain expressions for $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$ we use the δ-**method**.

The δ-method

The δ-method is developed to address the following problem

Problem

Suppose that X and Y are random variables, and that $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and $\sigma_{XY} = Cov(X, Y)$ are known. The problem is to find μ_Z and σ_Z^2, where $Z = f(X, Y)$.

Using the Taylor series expansion to the first order:

$$Z = f(X, Y) \approx f(\mu) + (X - \mu_X)\frac{\partial f}{\partial x}(\mu) + (Y - \mu_Y)\frac{\partial f}{\partial y}(\mu), \qquad \mu = (\mu_X, \mu_Y)$$

Therefore,

$$\boxed{\mu_Z \approx f(\mu)} \qquad \boxed{\sigma_Z^2 \approx \sigma_X^2 \left(\frac{\partial f}{\partial x}(\mu)\right)^2 + \sigma_Y^2 \left(\frac{\partial f}{\partial y}(\mu)\right)^2 + 2\sigma_{XY}\frac{\partial f}{\partial x}(\mu)\frac{\partial f}{\partial y}(\mu)}$$

To obtain a better approximation for μ_Z, we can use the Taylor series expansion to the second order.

Approximations of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$

Using the δ-method, we obtain

Theorem

The expectation and variance of R_n are given by

$$\boxed{\mathbb{E}[R_n] \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \sigma_{xy})} \tag{2}$$

$$\boxed{\mathbb{V}[R_n] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})} \tag{3}$$

In applications, population parameters $\mu_x, \sigma_x, \sigma_y, \sigma_{xy}$ are unknown. To compute the **estimated** values of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$, we use (2) and (3) together with

- $r \approx R_n \qquad \mu_x \approx \overline{X}_n$
- $\sigma_x^2 \approx \hat{\sigma}^2_{x,\text{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)^2$
- $\sigma_y^2 \approx \hat{\sigma}^2_{y,\text{unbiased}} = \frac{N-1}{Nn-N}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$
- $\sigma_{xy} \approx \frac{N-1}{Nn-N}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$

Stratified Random Sampling

In Stratified Random Sampling, a population is partitioned into strata, which are then independently sampled using simple random sampling.

If $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ is the sample drawn from the k^{th} stratum, then the estimate of μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)} \approx \mu,$$

where $\omega_k = N_k/N$ is the fraction of the population in the k^{th} stratum.

- $\overline{X}_n^*$ is an unbiased estimate of μ

$$\mathbb{E}[\overline{X}_n^*] = \mu$$

- The variance of $\overline{X}_n^*$ is

$$\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k} \left(1 - \frac{n_k - 1}{N_k - 1}\right) \approx \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}$$

Neyman (=Optimal) Allocation Scheme

Question:

Suppose that the resources of a survey allow only a total of n units to be sampled. How to choose $n_1, \dots, n_L$ to minimize $\mathbb{V}[\overline{X}^*_n]$ subject to constraint $\sum n_k = n$?

Optimization problem:

$$\mathbb{V}[\overline{X}^*_n] \to \min \qquad \text{s.t.} \sum_{k=1}^{L} n_k = n \tag{4}$$

Theorem

- *The sample sizes $n_1, \dots, n_L$ that solve the optimization problem (4) are given by*

$$\hat{n}_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j} \qquad k = 1, \dots, L$$

- *The variance of the optimal stratified estimate is*

$$\mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n} \left(\sum_{k=1}^{L} \omega_k \sigma_k \right)^2$$

Proportional Allocation

There are two main disadvantages of Neyman allocation:

1. Optimal allocations $\hat{n}_k$ depends on σ_k which generally will not be known
2. If a survey measures several values for each population member, then it is usually impossible to find an allocation that is simultaneously optimal for all values

A simple and popular alternative method of allocation is proportional allocation: to choose $n_1, \ldots, n_L$ such that

$$\boxed{\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}}$$

This holds if

$$\tilde{n}_k = n\frac{N_k}{N} = n\omega_k \qquad k = 1, \ldots, L \tag{5}$$

Theorem

The variance of $\overline{X}^*_{n,p}$ is given by

$$\mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k\sigma_k^2$$

Neyman vs Proportional and Simple vs Stratified

By definition, Neyman allocation is always better than proportional allocation.

Question: When is it substantially better?

$$\mathbb{V}[\overline{X}^*_{n,p}] - \mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k(\sigma_k - \bar{\sigma})^2, \qquad \bar{\sigma} = \sum_{k=1}^{L}\omega_k\sigma_k$$

- if the variances σ_k of the strata are all the same, then proportional allocation is as efficient as Neyman allocation, $\mathbb{V}[\overline{X}^*_{n,p}] = \mathbb{V}[\overline{X}^*_{n,opt}]$
- the more variable σ_k, the more efficient the Neyman allocation scheme

Question: What is more efficient: simple random sampling or stratified random sampling with proportional allocation?

$$\mathbb{V}[\overline{X}_n] - \mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k(\mu_k - \mu)^2$$

Thus, stratified random sampling with proportional allocation always gives a smaller variance than simple random sampling does (providing that the finite population correction is ignored, $(n-1)/(N-1) \approx 0$).

Math 408 - Mathematical Statistics

Lecture 23a. Fundamental Concepts of Modern Statistical Inference

March 29, 2013

Agenda

- Statistical Models
- Point Estimates
- Confidence Intervals
- Hypothesis Testing
- Summary

Statistical Inference

Statistical inference, or "learning", is the process of using data to infer the distribution that generated the data. The basic statistical inference problem is the following:

Basic Problem

We observe $X_1, \ldots, X_n \sim \pi$. We want to infer (or estimate, or learn) π or some features of π such as its mean.

Definition

A **statistical model** is a set of distributions or a set of densities $\mathcal{F}$.

- A **parametric model** is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
- A **nonparametric model** is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.

Examples of Statistical Models

Examples:

- If we assume that the data come from a normal distribution, then the model is

$$\mathcal{F} = \left\{ \pi(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \mu, \sigma^2 \in \mathbb{R} \right\}$$

  This is a two-parameter model. In $\pi(x|\mu,\sigma^2)$, x is a value of the random variable, whereas μ and σ^2 are parameters.

- A nonparametric model:

$$\mathcal{F}_{\mathrm{all}} = \{\text{all PDFs}\}$$

We will focus on parametric models.
In general, a parametric model takes the form

$$\boxed{\mathcal{F} = \{\pi(x|\theta), \;\; \theta \in \Theta\}}$$

where θ is an unknown parameter and Θ is the parameter space.

Remark: θ can be a vector, for instance, $\theta = (\mu, \sigma^2)$

Point Estimation

Given a parametric model, $\mathcal{F} = \{\pi(x|\theta), \ \ \theta \in \Theta\}$, the problem of inference is then to estimate (to learn) the parameter θ from the data.

Almost all problems in statistical inference can be identified as being one of three types: **point estimates**, **confidence intervals**, and **hypothesis testing**.

- Point Estimation refers to providing a single "best guess."
  Suppose $X_1, \ldots, X_n \sim \pi(x|\theta)$, where $\pi(x|\theta) \in \mathcal{F}$.
  A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = f(X_1, \ldots, X_n)$$

  Remember: θ is fixed but unknown, $\hat{\theta}_n$ is random since depends on $X_1, \ldots, X_n$. We say that $\hat{\theta}_n$ is unbiased if

$$\mathbb{E}[\hat{\theta}_n] = \theta$$

Confidence Intervals and Hypothesis Testing

- A $100(1-\alpha)\%$ Confidence Interval for a parameter θ is a random interval $I_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that

$$\mathbb{P}(\theta \in I_n) = 1 - \alpha$$

  In words: (a, b) traps θ with probability $1 - \alpha$.
  $(1 - \alpha)$ is called coverage of the confidence interval.
  In practice, $\alpha = 0.05$ is often used.

- In Hypothesis Testing, we start with some default theory, called a null hypothesis, and we ask if the data provide sufficient evidence to reject the theory. If not, we accept the null hypothesis.
  Example:
  $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$: n independent coin flips.
  We want to test if the coin is fair $\Rightarrow$ the null hypothesis $H_0 : p = 1/2$
  The alternative hypothesis is then: $H_1 : p \neq 1/2$
  It seems reasonable to reject H_0 if

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{2} \right| \quad \text{is large}$$

Summary

- A parametric model is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
  - General parametric model:

$$\mathcal{F} = \{\pi(x|\theta), \quad \theta \in \Theta\}$$

- A nonparametric model is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.
- Almost all problems in statistical inference can be identified as being one of three types:
  - Point Estimates
  - Confidence Intervals
  - Hypothesis Testing

Math 408 - Mathematical Statistics

Lecture 23b. The Method of Moments

March 29, 2013

Method of Moments: Problem Formulation

Suppose that

$$X_1, \ldots, X_n \sim \pi(x|\theta)$$

where $\theta \in \Theta$, and we want to estimate θ based on the data $X_1, \ldots, X_n$.

The first method for constructing parametric estimators that we will study is called the method of moments.

- The estimators produced by this method are not optimal, but that are often easy to compute.
- They are also useful as starting values for other methods that require iterative numerical routines.

Method of Moments

Recall that the k^{th} moment of a probability distribution $\pi(x|\theta)$ is

$$\mu_k(\theta) = \mathbb{E}_\theta[X^k]$$

where $\mathbb{E}_\theta$ denotes expectation with respect to $\pi(x|\theta)$, i.e.

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

If $X_1, \ldots, X_n$ are i.i.d from $\pi(x|\theta)$, then the k^{th} sample moment if defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

We can view $\hat{\mu}_k$ as an estimate of μ_k. Suppose that the parameter θ has k components:

$$\theta = (\theta_1, \ldots, \theta_k)$$

Method of Moments

Method of Moments

The **method of moments estimator** $\hat{\theta}$ *is defined to be the value of* θ *such that*

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \quad \ldots\ldots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases} \tag{1}$$

- System (1) is a system of k equations with k unknowns: $\theta_1, \ldots, \theta_k$
- The solutions of this system $\hat{\theta}$ is the method of moments estimate of the parameter θ.

Example 1: Bernoulli

- Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$.
  Find the method of moments estimate of the parameter p.

Example 2: Normal

- Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
  Find the method of moments estimates of μ and σ^2.

Consistency of the MoM estimator

Question: How good is the estimator $\hat{\theta}$ obtained by the method of moments?

Definition

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is **consistent** if

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

That is, for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0 \text{ as } n \to \infty$$

Theorem

The method of moments estimate is consistent.

Summary

- If $X_1, \ldots, X_n \sim \pi(x|\theta)$, then the method of moments estimate $\hat{\theta}$ of $\theta = (\theta_1, \ldots, \theta_k)$ is the solution of

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \quad \ldots\ldots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases}$$

  where
  - $\mu_k(\theta)$ is the k^{th} moment
  $$\mu_k(\theta) = \mathbb{E}_\theta[X^k]$$
  - $\hat{\mu}_k$ is the k^{th} sample moment
  $$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The method of moments estimate $\hat{\theta}$ is a consistent estimate of θ.

Math 408 - Mathematical Statistics

Lecture 24. The Method of Maximum Likelihood

April 1, 2013

Agenda

- The Likelihood Function
- Maximum Likelihood Estimate (MLE)
- Properties of MLE
- Summary

The Likelihood Function

The most common method for estimating parameters in a parametric model is the method of maximum likelihood.

Suppose $X_1, \ldots, X_n$ are i.i.d. from $\pi(x|\theta)$.

Definition

The **likelihood function** is defined by

$$\boxed{\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi(X_i|\theta)}$$

Important Remarks:

- The likelihood function is just the joint pdf/pmf of the data, except that we treat it as a function of the parameter θ.
- Thus, $\mathcal{L} : \Theta \to [0, \infty)$
- The likelihood function is not a density function: it is not true that $\mathcal{L}$ integrates to one, i.e $\int_\Theta \mathcal{L}(\theta) d\theta \neq 1$.

Maximum Likelihood Estimate

Definition

The **maximum likelihood estimate** (MLE) of θ, denoted $\hat{\theta}_{\mathrm{MLE}}$, is the value of θ that maximizes the likelihood $\mathcal{L}(\theta)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

$\hat{\theta}_{\mathrm{MLE}}$ makes the observed data $X_1, \ldots, X_n$ "most probable" or "most likely"

Important Remark:

Rather than maximizing the likelihood itself, it is often easier to maximize its natural logarithm (which is equivalent since the log is a monotonic function). The log-likelihood is

$$\boxed{l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \pi(X_i|\theta)}$$

Example: Bernoulli

- $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$. Find the MLE of p.
- Answer:

$$\hat{p}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

- In this example, $\hat{p}_{\mathrm{MLE}} = \hat{p}_{\mathrm{MoM}}$

Example: Normal

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Find the MLEs of μ and σ^2.
- Answer:

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Again, in this example, MLEs are the same as the MoM estimates.

Properties of MLE

Under certain conditions on the model

$$\mathcal{F} = \{\pi(x|\theta), \;\; \theta \in \Theta\}$$

(under some smoothness conditions of π), the MLE $\hat{\theta}_{\mathrm{MLE}}$ possesses many attractive properties that make it an appealing choice of estimate.

Main properties of the MLE:

- MLE is consistent:

  $$\hat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathbb{P}} \theta_0$$

  where θ_0 denotes the true value of θ.

- MLE is equivariant:
  if $\hat{\theta}_{\mathrm{MLE}}$ is the MLE of $\theta \quad \Rightarrow \quad f(\hat{\theta}_{\mathrm{MLE}})$ is the MLE of $f(\theta)$.

- MLE is asymptotically optimal: the MLE has the smallest variance for large sample sizes n.

Properties of MLE

Main properties of the MLE (cont):

- MLE is asymptotically Normal:

$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

  where

$$I(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\pi(X|\theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log\pi(x|\theta)\right)^2\pi(x|\theta)dx$$

  - $I(\theta)$ is called Fisher Information.

- MLE is asymptotically unbiased:

$$\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_{\mathrm{MLE}}] = \theta_0$$

Example: when MoM and MLE produce different estimates

- $X_1, \ldots, X_n \sim U(0, \theta)$. Find the MoM estimate and MLE of θ.
- Answer:

$$\hat{\theta}_{\mathrm{MoM}} = 2\overline{X}_n \qquad \hat{\theta}_{\mathrm{MLE}} = X_{(n)}$$

- In this example, the MLE and MoM estimate are different.

Summary

- The Likelihood Function:

$$\boxed{\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi(X_i|\theta)} \qquad X_1, \ldots, X_n \sim \pi(x|\theta)$$

- The Maximum Likelihood Estimate:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta) = \arg\max_{\theta \in \Theta} \log \mathcal{L}(\theta)$$

- MLE is consistent, equivariant, asymptotically optimal, asymptotically normal, and asymptotically unbiased.
- Examples: $\mathrm{Bernoulli}(p)$, $N(\mu, \sigma^2)$, and $U(0, \theta)$.

Math 408 - Mathematical Statistics

Lecture 26-27. Confidence Intervals from MLEs

April 5-8, 2013

Agenda

- Exact Method
  - Normal distribution $N(\mu, \sigma^2)$
- Approximate Method
  - $\mathrm{Bernoulli}(p)$
- Bootstrap Method
- Summary

Three Methods

Recall the definition of a confidence interval (see also Lectures 8,17,23):

Definition

A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a random interval calculated from the sample,

$$X_1, \ldots, X_n \sim \pi(x|\theta)$$

which contains θ with probability $1-\alpha$.

There are three methods for constructing confidence intervals using MLEs $\hat{\theta}_{\mathrm{MLE}}$:

- Exact Method
- Approximate Method
- Bootstrap Method

Exact Method. Example: Normal distribution $\mathcal{N}(\mu, \sigma^2)$

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then the MLEs for μ and σ^2 are (Lecture 24):

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- A confidence interval for μ is based on the following fact (Lecture 13-14):

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}$$

where S_n^2 is the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\hat{\sigma}^2_{\mathrm{MLE}}$

Result

A $100(1-\alpha)\%$ confidence interval for μ is

$$\hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}}\hat{\sigma}_{\mathrm{MLE}} t_{n-1}(\alpha/2)$$

where $t_{n-1}(\alpha)$ is the point beyond which the t-distribution with $(n-1)$ degrees of freedom has probability α.

Exact Method. Example: Normal distribution $N(\mu, \sigma^2)$

- A confidence interval for σ^2 is based on the following fact (Lecture 13-14):

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Result

A $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(1-\frac{\alpha}{2})}\right)$$

where $\chi^2_{n-1}(\alpha)$ is the point beyond which the χ^2-distribution with $(n-1)$ degrees of freedom has probability α.

Remark:

The main drawback of the exact method is that in practice the sampling distributions — like t_{n-1} and χ^2_{n-1} in our example — are not known.

Approximate Method

One of the most important properties of MLE is that it is asymptotically normal:

$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right), \qquad \text{as } n \to \infty$$

where $I(\theta_0)$ is Fisher information

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\pi(X|\theta)\right)^2\right]$$

Since the true value θ_0 is unknown, we will use $I(\hat{\theta}_{\mathrm{MLE}})$ instead of $I(\theta_0)$:

Result

An **approximate** $100(1-\alpha)\%$ *confidence interval for* θ_0 *is*

$$\hat{\theta}_{\mathrm{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{\mathrm{MLE}})}}$$

where z_α *is the point beyond which the standard normal distribution has probability* α*.*

Approximate Method. Example: $\mathrm{Bernoulli}(p)$

- Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$.
  Find an approximate confidence interval for p
- Answer:

$$\overline{X}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}$$

Bootstrap Method

Suppose $\hat{\theta}$ is an estimate of a parameter θ, the true unknown value of which is θ_0. $\hat{\theta}$ can be any estimate, not necessarily MLE,

$$X_1, \ldots, X_n \sim \pi(x|\theta) \qquad \hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

Define a new random variable

$$\Delta = \hat{\theta} - \theta_0$$

- Step 1: Assume (for the moment) that the distribution of Δ is known. Let (as before) q_α be the number such that $\mathbb{P}(\Delta > q_\alpha) = \alpha$.Then

$$\mathbb{P}(q_{1-\frac{\alpha}{2}} \leq \hat{\theta} - \theta_0 \leq q_{\frac{\alpha}{2}}) = 1 - \alpha$$

  And therefore a $100(1-\alpha)\%$ confidence interval for θ_0 is

$$\left(\hat{\theta} - q_{\frac{\alpha}{2}}, \hat{\theta} - q_{1-\frac{\alpha}{2}}\right)$$

  The problem is that the distribution of Δ is not known and, therefore, q_α are not known.

Bootstrap Method

- Step 2: Assume that the distribution of Δ is not known, but θ_0 is known. Then we can approximate the distribution of Δ as follows:

$$
\begin{aligned}
X_1^{(1)}, \ldots, X_n^{(1)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(1)} - \theta_0 = \Delta^{(1)} \\
X_1^{(2)}, \ldots, X_n^{(2)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(2)} - \theta_0 = \Delta^{(2)} \\
&\cdots\cdots \\
X_1^{(B)}, \ldots, X_n^{(B)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(B)} - \theta_0 = \Delta^{(B)}
\end{aligned}
$$

From these realizations $\Delta^{(1)}, \ldots, \Delta^{(B)}$ of Δ we can approximate the distribution of Δ by its empirical distribution, and, therefore, we can approximate q_α. The problem is that θ_0 is not known!

Bootstrap Method

- Step 3: **Bootstrap strategy**: Use $\hat{\theta}$ instead of θ_0.

$$
\begin{aligned}
X_1^{(1)},\dots,X_n^{(1)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(1)}-\hat{\theta} \approx \Delta^{(1)} \\
X_1^{(2)},\dots,X_n^{(2)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(2)}-\hat{\theta} \approx \Delta^{(2)} \\
&\cdots\cdots \\
X_1^{(B)},\dots,X_n^{(B)} \sim \pi(x|\theta_0) \quad &\rightsquigarrow \quad \hat{\theta}^{(B)}-\hat{\theta} \approx \Delta^{(B)}
\end{aligned}
$$

Distribution of Δ is approximated from realizations $\Delta^{(1)},\dots,\Delta^{(B)}$.

Remark:

$\hat{\theta}^{(i)}$ is the estimate of θ that is obtained from $X_1^{(i)},\dots,X_n^{(i)}$ by the same method (for example, MLE) as $\hat{\theta}$ was obtained from $X_1,\dots,X_n$.

Summary

- We considered three methods for constructing confidence intervals using MLEs: Exact Method, Approximate Method, Bootstrap Method
- Exact Method provides exact confidence intervals, but it is difficult to use in practice
  - Example: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu : \quad \hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}} \hat{\sigma}^2_{\mathrm{MLE}} t_{n-1}(\alpha/2)$$

$$\sigma^2 : \quad \left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(1-\frac{\alpha}{2})} \right)$$

- Approximate method provides an approximate confidence interval for θ_0, which is constructed using asymptotical properties of MLE:

$$\hat{\theta}_{\mathrm{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{\mathrm{MLE}})}}$$

- Bootstrap Method provides an approximate confidence interval. Bootstrap is the most popular method in practice since it is easy to implement.

Math 408 - Mathematical Statistics

Lecture 28. Efficiency and the Cramer-Rao Lower Bound

April 10, 2013

Agenda

- Mean Squared Error
- Cramer-Rao Inequality
- Example: Poisson Distribution
- Summary

Measure of Efficiency: Mean Squared Error

In most estimation problems, there are many possible estimates $\hat{\theta}$ of θ.
For example, the MoM estimate $\hat{\theta}_{\mathrm{MoM}}$ or the MLE estimate $\hat{\theta}_{\mathrm{MLE}}$.

Question: How would we choose which estimate to use?

Qualitatively, it is reasonable to choose that estimate whose distribution is most highly concentrated about the true parameter value θ_0. To make this idea work, we need to define a quantitative measure of such concentration.

Definition

The **mean squared error** of $\hat{\theta}$ as an estimate of θ_0 is

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]}$$

- The mean squared error can be also written as follows:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] + \underbrace{(\mathbb{E}(\hat{\theta}) - \theta_0)^2}_{\text{squared bias}}$$

- If $\hat{\theta}$ is unbiased, then $\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}]$.

Cramer-Rao Inequality

- Given two unbiased estimates, $\hat{\theta}$ and $\tilde{\theta}$, the **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\mathrm{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\mathbb{V}[\tilde{\theta}]}{\mathbb{V}[\hat{\theta}]}$$

- $\hat{\theta}$ is more efficient than $\tilde{\theta}$ $\Leftrightarrow$ $\mathrm{eff}(\hat{\theta}, \tilde{\theta}) > 1$
- In general, the mean squared error is a measure of efficiency of an estimate:

the smaller $\mathrm{MSE}(\hat{\theta})$, the better the estimate $\hat{\theta}$

Cramer-Rao Inequality

Let $X_1, \ldots, X_n$ be i.i.d. from $\pi(x|\theta)$. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be any unbiased estimate of a parameter θ whose true values is θ_0. Then, under smoothness assumptions on $\pi(x|\theta)$,

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}$$

Cramer-Rao Inequality

Cramer-Rao: $$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}}$$

Important Remarks:

- $\hat{\theta}$ can't have arbitrary small MSE
- The Cramer-Rao inequality gives a lower bound on the variance of any unbiased estimate.

Definition

An unbiased estimate whose variance achieves this lower bound is said to be **efficient**.

Recall that MLE is asymptotically Normal: $\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$

- Therefore, MLE is asymptotically efficient
- However, for a finite sample size n, MLE may not be efficient
- MLEs are not the only asymptotically efficient estimates.

Example: Poisson Distribution

Recall that the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events k occurring in a fixed interval of time if these events occur with a known average rate λ and independently of the time since the last event.

$$\mathbb{P}(X = k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad \mathbb{E}[X] = \lambda \qquad \mathbb{V}[X] = \lambda$$

Example

Let $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda)$.

- Find the MLE of λ
- Show that $\hat{\lambda}_{\mathrm{MLE}}$ is efficient.

- The theorem does not exclude the possibility that there is a biased estimator of λ that has a smaller MSE than $\hat{\lambda}_{\mathrm{MLE}}$

Summary

- Mean squared error is a measure of efficiency of an estimate

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]$$

- If $\hat{\theta}$ is unbiased, then

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}]$$

- Cramer-Rao Inequality:

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}}$$

- An unbiased estimate whose variance achieves this lower bound is said to be efficient
- Any MLE is asymptotically efficient (as $n \to \infty$)
- Example: if $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$, then $\hat{\lambda}_{\mathrm{MLE}}$ is efficient

Math 408 - Mathematical Statistics

Lecture 29-30. Testing Hypotheses: The Neyman-Pearson Paradigm

April 12-15, 2013

Agenda

- Example: Two Coins Tossing
- General Framework
- Type I Error and Type II Error
- Significance Level
- Power
- Neyman-Pearson Lemma
- Example: the likelihood ratio test for Gaussian variables
- The Concept of p-value
- Summary

Example: Two Coins Tossing

Suppose Bob has two coins:

- Coin "0" has probability of heads $p_0 = 0.5$
- Coin "1" has probability of heads $p_1 = 0.7$

Bob chooses one of the coins, tosses it $n = 10$ times and tells Alice the number of heads, but does not tell her whether it was coin 0 or coin 1.

On the basis of the number of heads, Alice has to decide which coin it was.
How should her decision rule be?

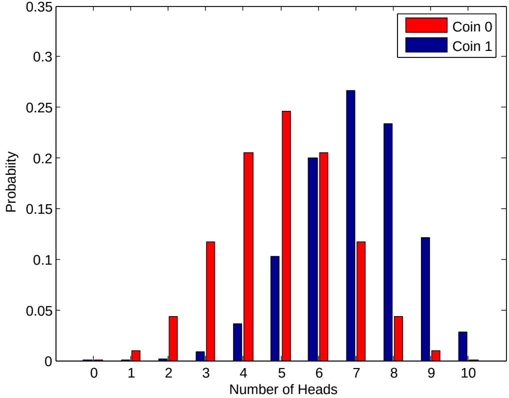
Let X denote the number of heads.

$$X \in \mathcal{X} = \{0, 1, 2, \ldots, 10\}$$

Then for each coin we can compute the probability that Bob got exactly x heads:

$$\mathbb{P}_i(X = x) = \binom{n}{x} p_i^x (1 - p_i)^{n-x}, \qquad i = 0, 1.$$

Example: Two Coins Tossing



Suppose that Bob observed 2 heads. Then $\frac{\mathbb{P}_0(X=2)}{\mathbb{P}_1(X=2)} \approx 30$, and, therefore, coin 0 was about 30 times more likely to produce this result than was coin 1.

On the other hand, if there were 8 heads, then $\frac{\mathbb{P}_0(X=8)}{\mathbb{P}_1(X=8)} \approx 0.19$, which would favor coin 1.

Hypothesis Testing

The example with two coins is an example of **hypothesis testing**:

- The Null Hypothesis H_0: Bob tossed coin 0
- The Alternative Hypothesis H_1: Bob tossed coin 1

Alice would accept H_0 if the likelihood ratio

$$\frac{\mathcal{L}(\text{Data}|\text{Coin } 0)}{\mathcal{L}(\text{Data}|\text{Coin } 1)} = \frac{\mathbb{P}_0(X = x)}{\mathbb{P}_1(X = x)} > 1$$

and she would reject H_0 if the likelihood ratio

$$\frac{\mathcal{L}(\text{Data}|\text{Coin } 0)}{\mathcal{L}(\text{Data}|\text{Coin } 1)} = \frac{\mathbb{P}_0(X = x)}{\mathbb{P}_1(X = x)} < 1$$

In this example, Alice would accept H_0 if

$$x \leq 6$$

and she would reject H_0 if

$$x > 6$$

Hypothesis Testing: General Framework

More formally, 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**.

Let X be data and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $\mathcal{R} \subset \mathcal{X}$ called the **rejection region**. If $X \in \mathcal{R}$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$X \in \mathcal{R} \Rightarrow \text{ reject } H_0$$
$$X \notin \mathcal{R} \Rightarrow \text{ accept } H_0$$

In the Two Coins Example,

- X is the number of heads
- $\mathcal{X}$ is $\{0, 1, 2, \dots, 10\}$
- $\mathcal{R}$ is $\{7, 8, 9, 10\}$

Hypothesis Testing: General Framework

Usually the rejection region $\mathcal{R}$ is of the form

$$\mathcal{R} = \{x \in \mathcal{X} : T(x) < c\}$$

where T is a **test statistic** and c is a **critical value**. The main problem in hypothesis testing is

to find an appropriate test statistic T and an appropriate cutoff value c

In the Two Coins Example,

- $T(x) = \frac{\mathbb{P}_0(X=x)}{\mathbb{P}_1(X=x)}$ is the likelihood ratio
- $c = 1$

Main Definitions

In hypothesis testing, there are two types of errors we can make:

- Rejecting H_0 when H_0 is true is called a **type I error**
- Accepting H_0 when H_1 is true is called a **type II error**

Definition

- The probability of a type I error is called the **significance level** of the test and is denoted by α

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(\text{Reject } H_0 | H_0)$$

- The probability of a type II error is denoted by β

$$\beta = \mathbb{P}(\text{type II error}) = \mathbb{P}(\text{Accept } H_0 | H_1)$$

- $(1 - \beta)$ is called the **power** of the test

$$\textbf{power} = 1 - \beta = 1 - \mathbb{P}(\text{Accept } H_0 | H_1) = \mathbb{P}(\text{Reject } H_0 | H_1)$$

Thus, the **power** of the test is the probability of rejecting H_0 when it is false.

Neyman-Pearson Lemma

Definition

- A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**.
- A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**.

The Neyman-Pearson Lemma shows that the test that is based on the likelihood ratio (as in the Two Coins Example) is optimal for simple hypotheses:

Neyman-Pearson Lemma

Suppose that H_0 and H_1 are simple hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Suppose that the **likelihood ratio test** *that rejects H_0 whenever the likelihood ratio is less than c,*

$$\text{Reject } H_0 \quad \Leftrightarrow \quad \frac{\mathcal{L}(Data|\theta_0)}{\mathcal{L}(Data|\theta_1)} < c$$

has significance level α_{LR}. Then any other test for which the significance level $\alpha \leq \alpha_{\mathrm{LR}}$ has power less than or equal to that of the likelihood ratio test

$$1 - \beta \leq 1 - \beta_{LR}$$

Example

Example

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. Consider two simple hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1 > \mu_0$$

Construct the likelihood ratio test with significance level α.

Answer:

$$\text{Reject } H_0 \quad \Leftrightarrow \quad \overline{X}_n > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Neyman-Pearson: this test is the most powerful test among all tests with significance level α.

The Concept of p-value

Reporting "reject H_0" or "accept H_0" is not very informative.
For example, if the test just reposts to reject H_0, this does not tell us how strong the evidence against H_0 is. This evidence is summarized in terms of **p-value**.

Definition

Suppose for every $\alpha \in (0, 1)$ we have a test of significance level α with rejection region $\mathcal{R}_\alpha$. Then, the p-value is the smallest significance level at which we can reject H_0:

$$p\text{-value} = \inf\{\alpha : X \in \mathcal{R}_\alpha\}$$

Informally, the p-value is a measure of the evidence against H_0:
the smaller the p-value, the stronger the evidence against H_0
Typically, researchers use the following evidence scale:

- p-value < 0.01: very string evidence against H_0
- $0.01 < p$-value < 0.05: strong evidence against H_0
- $0.05 < p$-value < 0.10: weak evidence against H_0
- p-value > 0.10: little or no evidence against H_0

Summary

- In general, we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and test

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

    - H_0 is called the null hypothesis
    - H_1 is called the alternative hypothesis
    - If $H_i : \theta = \theta_i$, then the hypothesis is called simple
- If X is data and $\mathcal{X}$ is the range of X, then we reject $H_0 \Leftrightarrow X \in \mathcal{R} \subset \mathcal{X}$.
    - Rejection region $\mathcal{R} = \{x : T(x) < c\}$
    - For the likelihood ratio test, $T(x) = \frac{\mathbb{P}(X=x|H_0)}{\mathbb{P}(X=x|H_1)}$
- Type I Error: Rejecting H_0 when H_0 is true
    - $\alpha = \mathbb{P}(\text{Reject } H_0|H_0)$ is called significance level (small α is good)
- Type II Error: Accepting H_0 when H_1 is true
    - $1 - \beta = 1 - \mathbb{P}(\text{Accept } H_0|H_1)$ is called power (large power is good)
- Neyman-Pearson Lemma: basing the test on the likelihood ratio is optimal.
- p-value summarizes the evidence against the null hypothesis, $p\text{-value} = \inf\{\alpha : X \in \mathcal{R}_\alpha\}$.

Math 408 - Mathematical Statistics

Lecture 31. Generalized Likelihood Ratio Tests

April 17, 2013

Generalization of the Likelihood Ratio Test

The Neyman-Pearson Lemma says that the likelihood ratio test is optimal for simple hypotheses.

Goal: to develop a generalization of this test for use in situations in which the hypotheses are not simple

- Generalized likelihood ratio tests are not generally optimal, but they perform reasonably well.
  - Often there are no optimal tests at all.
- Generalized likelihood ratio tests have wide utility.
  - They play the same role in testing as MLEs do in estimation

Generalized Likelihood Ratio Test

Let $X = (X_1, \dots, X_n)$ be data and let $\pi(x|\theta)$ be the joint density of the data. The likelihood function is then

$$\mathcal{L}(\theta) = \pi(X|\theta)$$

Suppose we wish to test

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

where Θ_0 and Θ_1 are two disjoint sets of the parameter space Θ, $\Theta = \Theta_0 \sqcup \Theta_1$.

- Based on the data, a measure of relative plausibility of the hypotheses is the ratio of their likelihoods.
- If the hypotheses are composite, each likelihood is evaluated at that value of θ that maximizes it.

This yields the **generalized likelihood ratio**:

$$\boxed{\Lambda^* = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta_1} \mathcal{L}(\theta)}}$$

Small values of Λ^* tend to discredit H_0.

Generalized Likelihood Ratio Test

For technical reasons, it is preferable to use the following statistic instead of Λ^*:

$$\boxed{\Lambda = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta} \mathcal{L}(\theta)}}$$

- Λ is called the likelihood ratio statistic.
- Note that

  $$\Lambda = \min\{\Lambda^*, 1\}$$

  Thus, small values of Λ^* correspond to small values of Λ.

The rejection region $\mathcal{R}$ for a generalized likelihood test has the following form:

$$\text{reject } H_0 \quad \Leftrightarrow \quad X \in \mathcal{R} = \{X : \Lambda(X) < \lambda\}$$

The threshold λ is chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha,$$

where α is the desired significance level of the test.

Example

Let $X_1, \ldots, X_n$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, where variance σ^2 is known. Consider testing the following hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0$$

Construct the generalized likelihood test with significance level α.

Answer:

$$\text{Reject } H_0 \quad \Leftrightarrow \quad \frac{\sqrt{n}|\overline{X}_n - \mu_0|}{\sigma} > z_{\frac{\alpha}{2}}$$

Distribution of $\Lambda(X)$

In order for the generalized likelihood ratio test to have the significance level α, the threshold λ must be chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha$$

If the distribution of $\Lambda(X)$ under H_0 is known, then we can determine λ.

- In the Example, $-2 \log \Lambda(X) \sim \chi_1^2$.

Generally, the distribution of Λ is not of a simple form, but in many situations the following theorem provides the basis for an approximation of the distribution.

Theorem

Under smoothness conditions on $\pi(x|\theta)$, the null distribution of $-2\log\Lambda(X)$ (i.e. distribution under H_0) tends to a χ_d^2 as the sample size $n \to \infty$, where

$$d = \dim\Theta - \dim\Theta_0,$$

where $\dim\Theta$ and $\dim\Theta_0$ are the numbers of free parameters in Θ and Θ_0.

- In the Example, $\dim\Theta = 1$ and $\dim\Theta_0 = 0$.

Summary

- Generalized likelihood ratio tests are used when the hypothesis are composite
  - They are not generally optimal, but perform reasonably well.
  - They play the same role in testing as MLEs do in estimation.
- The rejection region $\mathcal{R}$ for a generalized likelihood test has the following form:

$$\text{reject } H_0 \quad \Leftrightarrow \quad X \in \mathcal{R} = \{X : \Lambda(X) < \lambda\}$$

  - Λ is the likelihood ratio statistic,

$$\Lambda = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta} \mathcal{L}(\theta)}$$

  - The threshold λ is chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha,$$

    where α is the desired significance level of the test.
- As sample size $n \to \infty$, the null distribution of $-2\log\Lambda(X)$ tends to a χ^2_d, where

$$d = \dim\Theta - \dim\Theta_0$$

Math 408 - Mathematical Statistics

Lecture 32-33. Pearson's χ^2 Test For Multinomial Data

April 19-22, 2013

Agenda

- Multinomial Distribution and its Properties
- Construction the GLRT for Multinomial Data
- The MLE for Parameters of the Multinomial Distribution
- The GLRT with Significance Level α
- Pearson's χ^2 Test
- Asymptotic Equivalence of the GLRT and the Pearson's Test
- Example: Mendel's Peas
- Summary

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution.

Consider drawing a ball from a box which has balls with k different colors labeled color 1, color 2,..., color k. Let $p = (p_1, \dots, p_k)$, where p_i is the probability of drawing a ball of color i,

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1$$

Draw n times (independent draws with replacement) and let $X = (X_1, \dots, X_k)$, where X_i is the number of times that color i appeared.

$$\sum_{i=1}^{k} X_i = n$$

We say that X has a $\mathrm{Multinomial}(n, p)$ distribution.

Application: Multinomial distributions are useful when a "success-failure" description is insufficient to understand a system. Multinomial distributions are relevant to situations where there are more than two possible outcomes. For example, temperature = high, med, low.

Properties of the Multinomial Distribution

$$X \sim \mathrm{Multinomial}(n, p)$$

- n is the number of trials
- k is the number of possible outcomes
- $p = (p_1, \ldots, p_k)$, where p_i is the probability of observing outcome i
- $X = (X_1, \ldots, X_k)$, where X_i is the number of occurrences of outcome i

Theorem

- *The probability mass function of X is*

$$\pi_X(x|n,p) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

- *The marginal distribution of X_i is* $\mathrm{Binomial}(n, p_i)$
- *The mean and covariance matrix of X are*

$$\mathbb{E}[X] = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix} \qquad \mathbb{V}[X] = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \ldots & -np_1p_k \\ -np_1p_2 & np_2(1-p_2) & \ldots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_2 & -np_2p_k & \ldots & np_k(1-p_k) \end{pmatrix}$$

Constructing the GLRT

Suppose that $X \sim \mathrm{Multinomial}(n, p)$, where p is unknown, and we want to test

$$H_0 : (p_1, \ldots, p_k) = (\tilde{p}_1, \ldots, \tilde{p}_k) \equiv \tilde{p} \qquad \text{v.s.} \qquad H_1 : (p_1, \ldots, p_k) \neq (\tilde{p}_1, \ldots, \tilde{p}_k)$$

To construct the generalized likelihood ratio test, first, we need to determine the likelihood function $\mathcal{L}(p)$. In this case:

$$\mathcal{L}(p_1, \ldots, p_k) = \pi_X(X|n, p) = \frac{n!}{X_1! \ldots X_k!} p_1^{X_1} \ldots p_k^{X_k}$$

The likelihood ratio statistic is

$$\Lambda = \frac{\max_{p \in \Theta_0} \mathcal{L}(p)}{\max_{p \in \Theta} \mathcal{L}(p)} = \frac{\mathcal{L}(\tilde{p})}{\mathcal{L}(\hat{p}_{MLE})}$$

- $\Theta_0 = \{p : p = \tilde{p}\}$, $\dim \Theta_0 = 0$
- $\Theta = \{p : \sum_{i=1}^{k} p_i = 1\}$, $\dim \Theta = k - 1$

Thus, to proceed, we need to find the MLE of p.

The MLE of p and the GLRT with level α

Theorem

Let $X \sim \mathrm{Multinomial}(n, p)$. *The maximum likelihood estimator of* p *is*

$$\hat{p}_{MLE} = \begin{pmatrix} \frac{X_1}{n} \\ \vdots \\ \frac{X_k}{n} \end{pmatrix} = \frac{X}{n}$$

Therefore, the likelihood ratio statistic is

$$\Lambda = \prod_{i=1}^{k} \left(\frac{n\tilde{p}_i}{X_i}\right)^{X_i}$$

and

$$-2\log\Lambda = 2\sum_{i=1}^{k} X_i \log\left(\frac{X_i}{n\tilde{p}_i}\right) \ \dot{\sim}\ \chi^2_{k-1}, \quad \text{when } n \to \infty$$

The GLRT with significance level α rejects H_0 if and only if

$$\boxed{2\sum_{i=1}^{k} X_i \log\left(\frac{X_i}{n\tilde{p}_i}\right) > \chi^2_{k-1}(\alpha)}$$

Pearson's χ^2 Test

In practice, the Pearson's χ^2 test is often used. The test is based on the following statistic:

$$T = \sum_{i=1}^{k} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- $O_i = X_i$ is the observed data
- $E_i = \mathbb{E}[X_i] = n\tilde{p}_i$ is the expected value of X_i under H_0
- T is called the **Pearson's χ^2 statistic**

The Pearson's χ^2 statistic and $-2\log\Lambda$ are asymptotically equivalent under H_0

Theorem

- *Under H_0, $T \xrightarrow{\mathcal{D}} \chi^2_{k-1}$.*
- *Pearson's test: reject H_0 if $T > \chi^2_{k-1}(\alpha)$ has asymptotic significance level α.*
- *The p-value is $\mathbb{P}(\xi > t)$, where $\xi \sim \chi^2_{k-1}$ and t is the observed value of T.*

Remark: Pearson's test has been more commonly used than the GLRT, because it is easier to calculate (especially without a computer!)

Mendel's Peas

Example

Mendel bred peas with round yellow seeds and wrinkled green seeds.
There are four types of progeny:

- round yellow, wrinkled yellow, round green, wrinkled green.

The number of each type is multinomial with probability (p_1, p_2, p_3, p_4).
According to Mendel's theory:

$$H_0 : (p_1, p_2, p_3, p_4) = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right) \equiv \tilde{p}$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$.
Question: Based on these data, should we accept or reject the Mendel's theory?
Solution:

- The observed value of Pearson's χ^2 statistic is $t = \sum_{i=1}^{4} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} = 0.47$
- Let $\alpha = 0.05$. Then $\chi_3^2(\alpha) = F_{\chi_3^2}^{-1}(1-\alpha) \approx 7.8$.
- Since $T < \chi_3^2(\alpha)$, we accept H_0.
- The p-value is p-value $= \mathbb{P}(\xi > 0.47) = 1 - F_{\chi_3^2}(0.47) \approx 0.92$.
- No evidence against Mendel's theory.

Summary

- Multinomial distribution: $X \sim \mathrm{Multinomial}(n, p)$
  - The probability mass function of X is

$$\pi_X(x|n, p) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

  - The marginal distribution of X_i is $\mathrm{Binomial}(n, p_i)$
  - The maximum likelihood estimator of p is $\hat{p}_{MLE} = X/n$
- Suppose that $X \sim \mathrm{Multinomial}(n, p)$, p is unknown, and we want to test

$$H_0 : (p_1, \dots, p_k) = (\tilde{p}_1, \dots, \tilde{p}_k) \equiv \tilde{p} \quad \text{v.s.} \quad H_1 : (p_1, \dots, p_k) \neq (\tilde{p}_1, \dots, \tilde{p}_k)$$

  - GLRT with significance level α rejects H_0 if $\boxed{2\sum_{i=1}^{k} X_i \log\left(\frac{X_i}{n\tilde{p}_i}\right) > \chi^2_{k-1}(\alpha)}$
  - Pearson's test: reject H_0 if $\boxed{T = \sum_{i=1}^{k} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} > \chi^2_{k-1}(\alpha)}$
    - ⋆ Under H_0, the Pearson's χ^2 statistic $T \xrightarrow{\mathcal{D}} \chi^2_{k-1}$.
    - ⋆ Pearson's test has asymptotic significance level α.
    - ⋆ The p-value is $\mathbb{P}(\xi > t)$, where $\xi \sim \chi^2_{k-1}$ and t is the observed value of T.

Math 408 - Mathematical Statistics

Lecture 34. Summarizing Data

April 24, 2013

Agenda

- Methods Based on the CDF
  - The Empirical CDF
    - Example: Data from Uniform Distribution
    - Example: Data from Normal Distribution
  - Statistical Properties of the eCDF
  - The Survival Function
    - Example: Data from Exponential Distribution
  - The Hazard Function
    - Example: The Hazard Function for the Exponential Distribution
- Summary

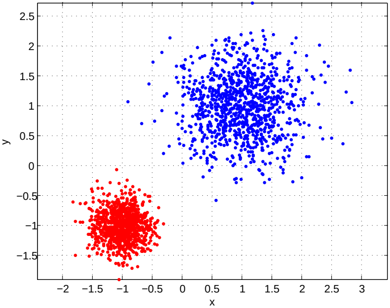
Describing Data

In the next few Lectures we will discuss methods for describing and summarizing data that are in the form of one or more samples. These methods are useful for revealing the structure of data that are initially in the form of numbers.

Example: the arithmetic mean $\overline{x} = (x_1 + \ldots + x_n)/n$ is often used as a summary of a collection of numbers $x_1, \ldots, x_n$: it indicates a "typical value".

Example:

- $x = (1.5147, 1.7223, 1.063, 1.4916, \ldots)$
- $y = (0.7353, 0.0781, 0.276, 1.5666, \ldots)$



Empirical CDF

Suppose that $x_1, \ldots, x_n$ is a batch of numbers.

<u>Remark:</u> We use the word

- "sample" when $X_1, \ldots, X_n$ is a collection of random variables.
- "batch" when $x_1, \ldots, x_n$ are fixed numbers (realization of sample).

Definition

The **empirical cumulative distribution function** (eCDF) is defined as

$$\boxed{F_n(x) = \frac{1}{n}(\# x_i \leq x)}$$

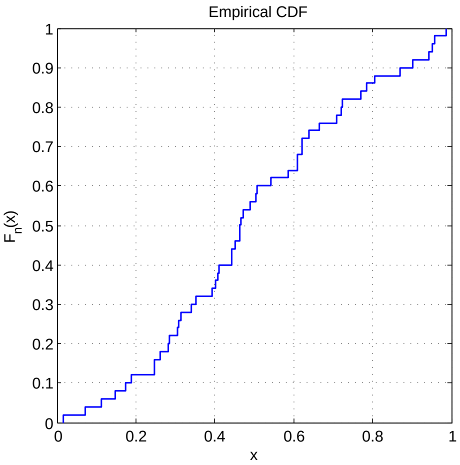
Denote the ordered batch of numbers by $x_{(1)}, \ldots, x_{(n)}$.

- If $x < x_{(1)}$, then $F_n(x) = 0$
- If $x_{(1)} \leq x < x_{(2)}$, then $F_n(x) = 1/n$
- If $x_{(k)} \leq x < x_{(k+1)}$, then $F_n(x) = k/n$

The eCDF is the "data analogue" of the CDF of a random variable

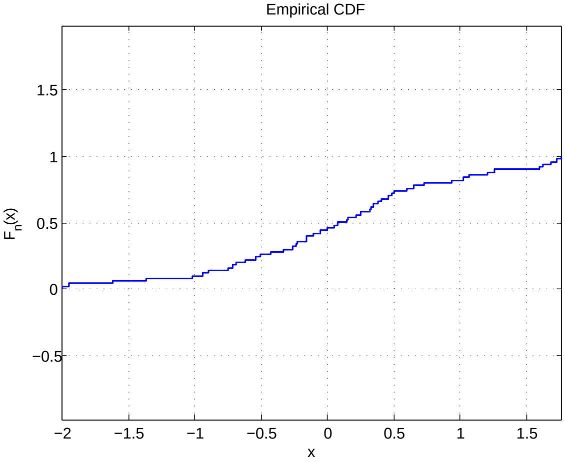
Example: Data from Uniform Distribution

- Let $(X_1, \ldots, X_n) \sim U[0, 1]$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (0.24733, 0.3527, 0.18786, 0.49064, \ldots)$



Example: Data from Normal Distribution

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (-0.23573, 0.45952, -0.93808, -0.62162, \ldots)$



Statistical Properties of the eCDF

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution F.
Then the eCDF can be written as follows:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

where

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}$$

The random variables $I_{(-\infty,x)}(X_1), \ldots, I_{(-\infty,x)}(X_n)$ are independent Bernoulli random variables:

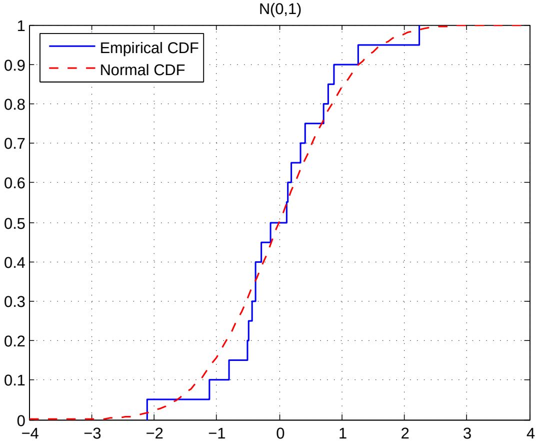
$$I_{(-\infty,x)}(X_i) = \begin{cases} 1, & \text{with probability } F(x) \\ 0, & \text{with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable: $nF_n(x) \sim \mathrm{Bin}(n, F(x))$

- $\mathbb{E}[F_n(x)] = F(x)$
- $\mathbb{V}[F_n(x)] = \frac{1}{n}F(x)(1 - F(x))$
- $\mathbb{V}[F_n(x)] \to 0, \;\text{ as }\; n \to \infty$

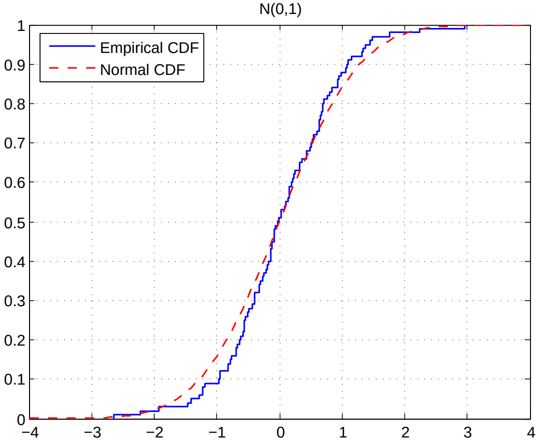
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 20$



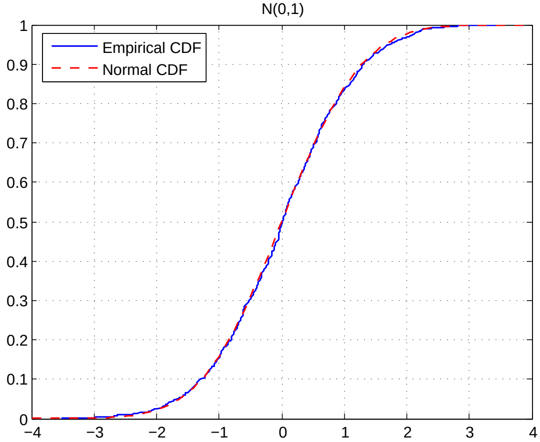
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0,1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 100$



Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 1000$



The Survival Function

The **survival function** is equivalent to the CDF and is defined as

$$\boxed{S(t) = \mathbb{P}(T > t) = 1 - F(t)}$$

In applications where the data consists of times until failure or death (and are thus nonnegative), it is often customary to work with the survival function rather than the CDF, although the two give equivalent information.
Data of this type occur in

- medical studies
- reliability studies

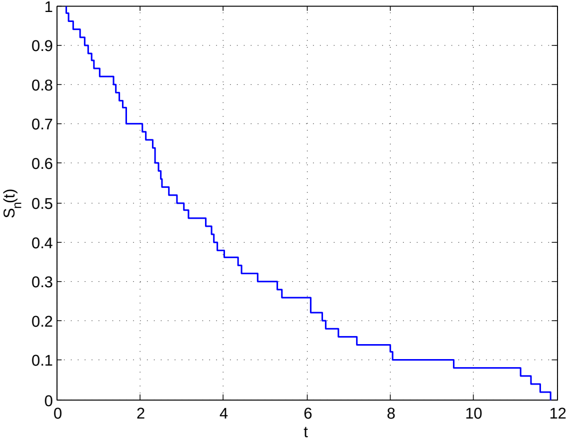
$$\boxed{S(t) = \text{ Probability that the lifetime will be longer than } t}$$

The data analogue of $S(t)$ is the **empirical survival function**:

$$\boxed{S_n(t) = 1 - F_n(t)}$$

Example: Data from Exponential Distribution

- Let $(X_1, \ldots, X_n) \sim \mathrm{Exp}(\beta)$, $\beta = 5$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (4.4356, 1.684, 11.376, 4.8357, \ldots)$



The Hazard Function

Let T is a random variable (time) with the CDF F and PDF f.

Definition

The **hazard function** is defined as

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}}$$

- The **hazard function** may be interpreted as the instantaneous death rate for individuals who have survived up to a given time: if an individual is alive at time t, the probability that individual will die in the time interval $(t, t+\epsilon)$ is

$$\mathbb{P}(t \leq T \leq t + \epsilon | T \geq t) \approx \frac{\epsilon f(t)}{1 - F(t)}$$

- If T is the lifetime of a manufactured component, it maybe natural to think of $h(t)$ as the age-specific failure rate. It may also be expressed as

$$h(t) = -\frac{d}{dt} \log S(t)$$

Example: Hazard Function for the Exponential Distribution

Let $T \sim \mathrm{Exp}(\beta)$, then

- $f(t) = \frac{1}{\beta} e^{-t/\beta}$
- $F(t) = 1 - e^{-t/\beta}$
- $S(t) = e^{-t/\beta}$
- $h(t) = \frac{1}{\beta}$

The instantaneous death rate is constant.

If the exponential distribution were used as a model for the lifetime of a component, it would imply that the probability of the component failing did not depend on its age.

Typically, a hazard function is U-shaped:

- the rate of failure is high for very new components because of flaws in the manufacturing process that show up very quickly,
- the rate of failure is relatively low for components of intermediate age,
- the rate of failure increases for older components as they wear out.

Summary

- The empirical cumulative distribution function (eCDF) is

$$\boxed{F_n(x) = \frac{1}{n}(\#x_i \le x)}$$

- The survival function is equivalent to the CDF and is defined as

$$\boxed{S(t) = \mathbb{P}(T > t) = 1 - F(t)}$$

- The data analogue of $S(t)$ is the empirical survival function:

$$\boxed{S_n(t) = 1 - F_n(t)}$$

- The hazard function is

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}}$$

  - may be interpreted as the instantaneous death rate for individuals who have survived up to a given time

Math 408 - Mathematical Statistics

Lecture 35. Summarizing Data - II

April 26, 2012

Agenda

- Quantile-Quantile Plots
- Histograms
- Kernel Probability Density Estimate
- Summary

Quantile-Quantile Plots

Quantile-Quantile (Q-Q) plots are used for comparing two probability distributions.

Suppose that X is a continuous random variable with a strictly increasing CDF F.

Definition

The p^{th} **quantile** of F is that value x_p such that

$$F(x_p) = p \quad \text{or} \quad \boxed{x_p = F^{-1}(p)}$$

Suppose we want to compare two CDF: F and G.

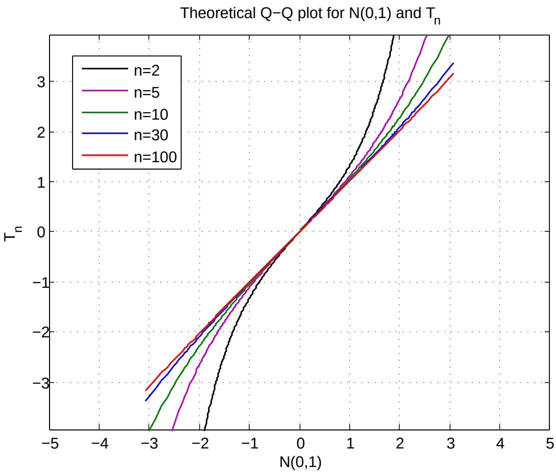
Definition

The **theoretical Q-Q plot** is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$, that is the graph $[F^{-1}(p), G^{-1}(p)]$ for $p \in (0, 1)$.

- If the two CDFs are identical, the theoretical Q-Q plot will be the line $y = x$.

Example of a Theoretical Q-Q plot

- $F = \mathcal{N}(0,1)$
- $G = T_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi_n^2/n}}$, t-distribution with n degrees of freedom.
- We know that $T_n \to \mathcal{N}(0,1)$ as $n \to \infty$.



Properties Q-Q plots

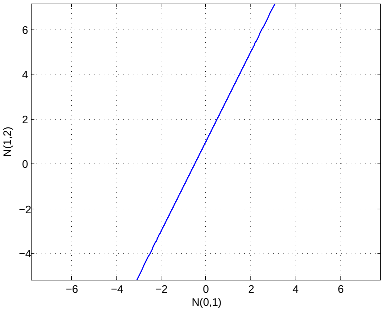
Theorem

If $G(x) = F(\frac{x-\mu}{\sigma})$ *for some constants* μ *and* $\sigma \neq 0$*, then*

$$\boxed{y_p = \mu + \sigma x_p}$$

- Thus, if two distributions differ only in location and/or scale, the theoretical Q-Q plot will be a straight line with slope σ and intercept μ.

Example: Let $F = \mathcal{N}(0,1)$ and $G = \mathcal{N}(1,2)$, then $G(x) = F(\frac{x-1}{\sqrt{2}})$.



Empirical Q-Q plots

In practice, a typical scenario is the following:

- $F(x) = F_0(x)$ is a specified CDF (e.g. normal) which is a theoretical model for data $X_1, \ldots, X_n$.
- $G(x)$ is the empirical CDF for $x_1, \ldots, x_n$, a realization of $X_1, \ldots, X_n$ (actually observed data).
- We want to compare the model $F(x)$ with the observation $G(x)$.

Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered batch. Then

Definition

The **empirical Q-Q plot** is the plot of $F_0^{-1}(i/n)$ on the horizonal axis versus $G^{-1}(i/n) = x_{(i)}$ on the vertical axis, for $i = 1, \ldots, n$.

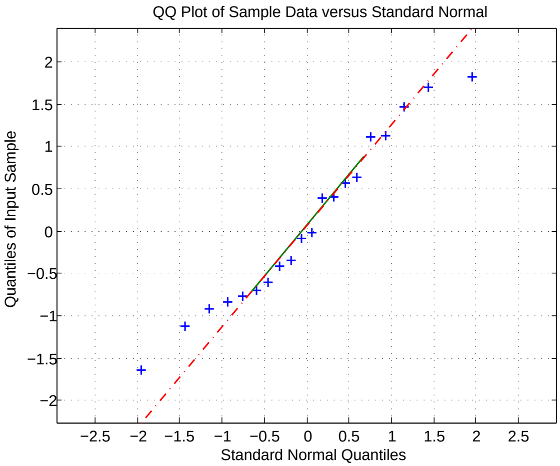
Remarks:

- The quantities $p_i = i/n$ are called plotting positions
- At $i = n$, there is a technical problem since $F_0^{-1}(1) = \infty$.
- Many software packages graph the following as the empirical Q-Q plot:

$$\left\{ \left(F_0^{-1}\left(\frac{i - 0.375}{n + 0.25}\right), x_{(i)} \right) \right\}, \qquad i = 1, \ldots, n$$

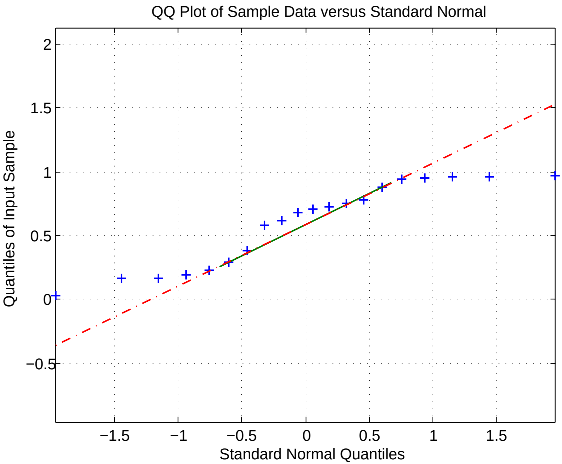
Example of an Empirical Q-Q plot

- $F_0 = \mathcal{N}(0,1)$, a model.
- $X_1, \ldots, X_{20} \sim \mathcal{N}(0,1)$.



Example of an Empirical Q-Q plot

- $F_0 = \mathcal{N}(0,1)$, a model.
- $X_1, \ldots, X_{20} \sim U[0,1]$.



Histograms

Histogram displays the shape of the distribution of data values.

Histograms are constructed in the following way:

1. The range of data $x_1, \ldots, x_n$ is divided into several intervals, called bins
2. The number of the observations falling in each bin is then plotted.

Remarks:

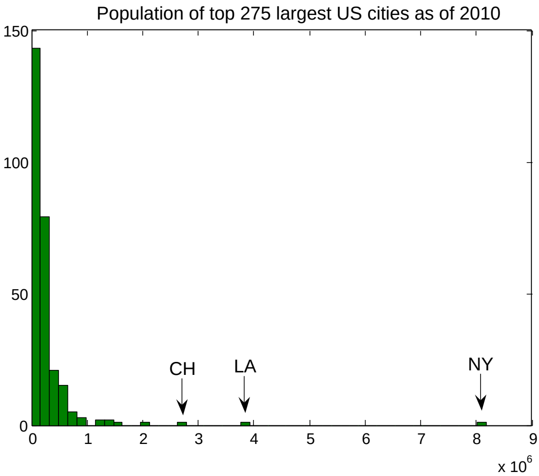
- The total area of the histogram is equal to the sample size n.
- A histogram may also be normalized displaying the proportion of observations falling in each bin. In this case, the area under the histogram is 1.

Applications:

- Histograms are frequently used to display data for which there is no assumption of any probability model. For example, populations of US cities.
- If the data are modeled as a random samples from some continuous distribution, then the normalized histogram may be also viewed as an estimate of the PDF.

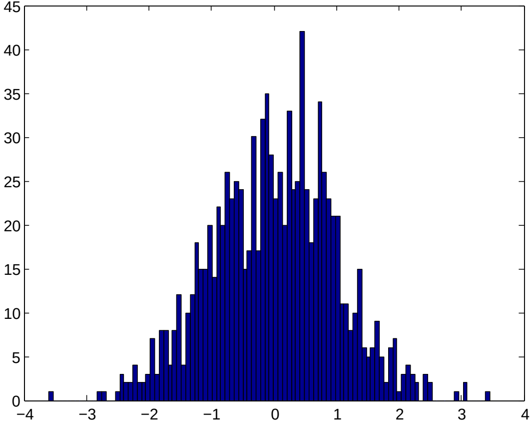
Example: Populations of US Cities

- Data $x_1, \ldots, x_{275}$ are populations of the top 275 largest US cities.
- Data source: `wikipedia.org`
- Number of bins: 50



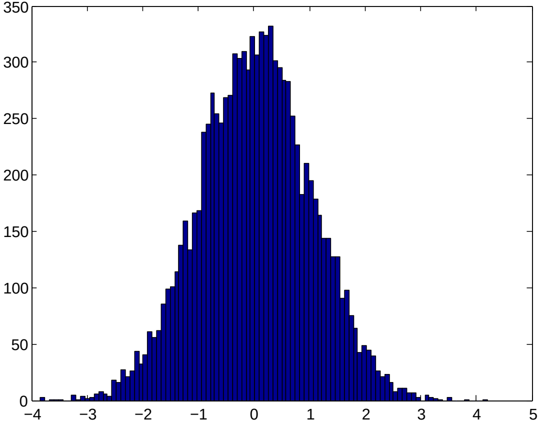
Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^3$
- Number of bins: 100



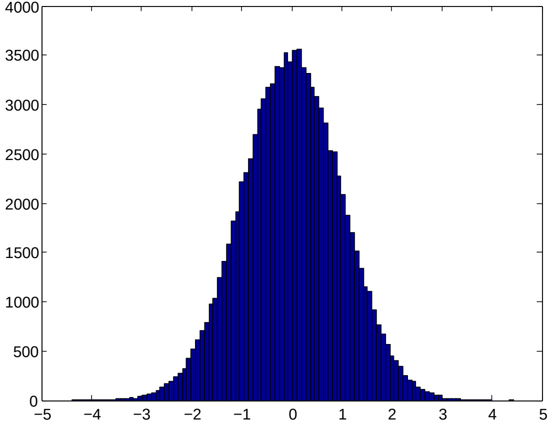
Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^4$
- Number of bins: 100



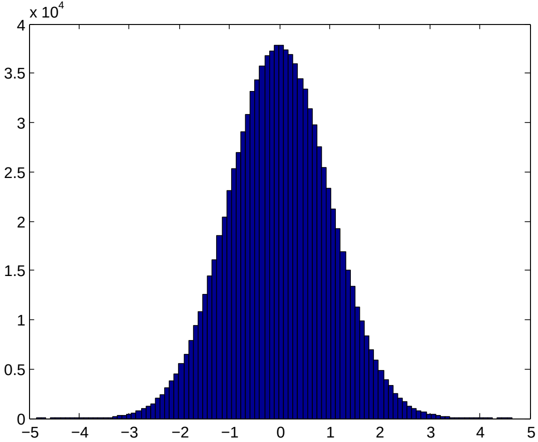
Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^5$
- Number of bins: 100



Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^6$
- Number of bins: 100



Kernel Probability Density Estimate

The main drawback of estimating PDFs by histograms is that these estimates are not smooth. A smooth probability density estimate can be constructed in the following way. Let $w(x)$ be a nonnegative, symmetric weight function, centered at zero and integrating to 1. For example, $w(x) = \mathcal{N}(x|0,1)$. The function

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

is a re-scaled version of $w(x)$.

- As $h \to 0$, $w_h(x)$ becomes more concentrated and peaked about zero.
- As $h \to \infty$, $w_h(x)$ becomes more spread out and flatter.
- If $w(x) = \mathcal{N}(x|0,1)$, then $w_h(x) = \mathcal{N}(x|0,h^2)$

Definition

If $X_1, \ldots, X_n \sim \pi$, then an estimate of π is

$$\boxed{\pi_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h(x - X_i)}$$

This estimate is called a **kernel probability density estimate**.

Kernel Probability Density Estimate

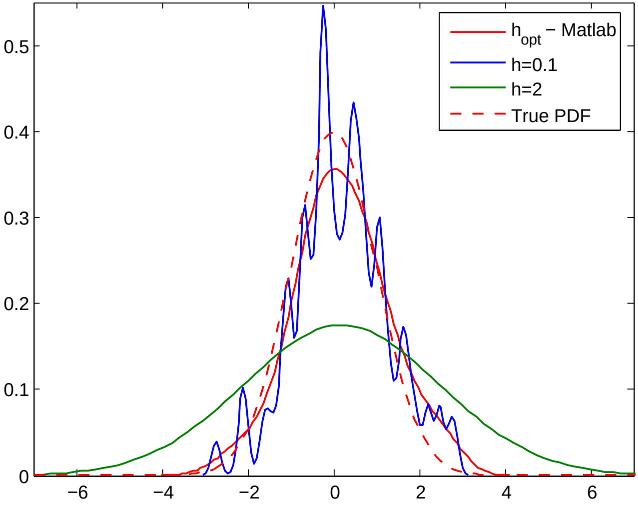
$$\boxed{\pi_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i)}$$

Remarks:

- $\pi_h(x)$ consists of the superposition of "hills" centered on the observations.
- If $w(x) = \mathcal{N}(x|0,1)$, then $w_h(x - X_i) = \mathcal{N}(x|X_i, h^2)$.
- The parameter h is called the bandwidth. It controls the smoothness of $\pi_h(x)$ and corresponds to the bin width of the histogram:
  - if h is too small, then $\pi_h(x)$ is too rough,
  - if h is too large, then the shape of $\pi_h(x)$ is smeared out too much.

Example

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 100$
- $w(x) = \mathcal{N}(x|0,1) \quad \Rightarrow \quad w_h(x - X_i) = \mathcal{N}(x|X_i, h^2)$.



Summary

- Quantile-Quantile (Q-Q) plots are used for comparing two distributions
  - The p^{th} quantile x_p of the CDF F is $x_p = F^{-1}(p)$
  - The theoretical Q-Q plot is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$.
  - If $F = G$, then the theoretical Q-Q plot will be the line $y = x$.
  - If $G(x) = F(\frac{x-\mu}{\sigma})$ for some constants μ and $\sigma \neq 0$, then $y_p = \mu + \sigma x_p$.
  - The empirical Q-Q plot is the plot of $F_0^{-1}(i/n)$ on the horizonal axis versus $x_{(i)}$ on the vertical axis.
- Histogram displays the shape of the distribution of data values.
  - Histograms are frequently used to display data for which there is no assumption of any probability model.
  - Normalized histogram may be also viewed as a non-smooth estimate of PDF.
- Kernel Probability Density Estimate: If $X_1, \ldots, X_n \sim \pi$, then an estimate of π is

$$\pi_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i)$$

  - If $w(x) = \mathcal{N}(x|0,1)$, then $w_h(x - X_i) = \mathcal{N}(x|X_i, h^2)$
  - h is the bandwidth.

Math 408 - Mathematical Statistics

Lecture 36. Summarizing Data - III

April 29, 2013

Agenda

- Measures of Location
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates
- Measures of Dispersion
  - Sample Standard Deviation
  - Interquartile Range (IQR)
  - Median Absolute Deviation (MAD)
- Boxplots
- Summary

Measures of Location

In Lectures 34 and 35, we discussed data analogues of the CDFs and PDFs, which convey visual information about the shape of the distribution of the data.

Next Goal: to discuss simple numerical summaries of data that are useful when there is not enough data for construction of an eCDF, or when a more concise summary is needed.

- A **measure of location** is a measure of the center of a batch of numbers.
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates

Example: If the numbers result from different measurement of the same quantity, a measure of location is often used in the hope that it is more accurate than any single measurement.

The Arithmetic Mean

The most commonly used measure of location is the arithmetic mean,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

A common statistical model for the variability of a measurement process is the following:

$$x_i = \mu + \varepsilon_i$$

- x_i is the value of the i^{th} measurement
- μ is the true value of the quantity
- ε_i is the random error, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

The arithmetic mean is then:

$$\overline{x} = \mu + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i, \qquad \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{n})$$

The Median

The main drawback of the arithmetic mean is it is sensitive to outliers. If fact, by changing a single number, the arithmetic mean of a batch of numbers can be made arbitrary large or small. For this reason, measures of location that are robust, or insensitive to outliers, are important.

Definition

If the batch size is an odd number, $x_1, \ldots, x_{2n-1}$, then the **median** $\tilde{x}$ is defined to be the middle value of the ordered batch values:

$$x_1, \ldots, x_{2n-1} \quad \rightsquigarrow \quad x_{(1)} < \ldots < x_{(2n-1)}, \quad \boxed{\tilde{x} = x_{(n)}}$$

Important Remark:

Moving the extreme observations does not affect the sample median at all, so the median is quite robust.

The Trimmed Mean

Another simple and robust measure of location is the **trimmed mean** or **truncated mean**.

Definition

The $100\alpha\%$ trimmed mean is defined as follows:

1. Order the data: $x_1, \dots, x_n \quad \rightsquigarrow \quad x_{(1)} < \dots < x_{(n)}$
2. Discard the lowest $100\alpha\%$ and the highest $100\alpha\%$
3. Take the arithmetic mean of the remaining data:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \dots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

where $[s]$ denotes the greatest integer less than or equal to s.

Remarks:

- It is generally recommended to use $\alpha \in [0.1, 0.2]$.
- Median can be considered as a 50% trimmed mean.

M Estimates

Let $x_1, \ldots, x_n$ be a batch of numbers. It is easy to show that

- The mean
$$\overline{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} (x_i - y)^2$$
Outliers have a great effect on mean, since the deviation of y from x_i is measured by the square of their difference.
- The median
$$\tilde{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} |x_i - y|$$
Here, large deviations are not weighted as heavily, that is exactly why the median is robust.

In general, consider the following function:

$$f(y) = \sum_{i=1}^{n} \Psi(x_i, y),$$

where Ψ is called the weight function. **M estimate** is the minimizer of f:

$$y^* = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$$

Measures of Dispersion

A measure of dispersion, or scale, gives a numerical characteristic of the "scatteredness" of a batch of numbers. The most commonly used measure is the sample standard deviation s, which is the square root of the sample variance,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Q: Why $\frac{1}{n-1}$ instead of $\frac{1}{n}$?

A: s^2 is an unbiased estimate of the population variance σ^2. If n is large, then it makes little difference whether $\frac{1}{n-1}$ or $\frac{1}{n}$ is used.

Like the mean, the standard deviation s is sensitive to outliers.

Measures of Dispersion

Two simple robust measures of dispersion are the interquartile range (IQR) and the median absolute deviation (MAD).

- **IQR** is the difference between the two sample quartiles:

$$\mathrm{IQR} = Q_3 - Q_1$$

  - Q_1 is the first (lower) quartile, splits lowest 25% of batch
  - $Q_2 = \tilde{x}$, cuts batch in half
  - Q_3 is the third (upper) quartile, splits highest 75% of batch

  How to compute the quartile values (one possible method):

  1. Find the median. It divides the ordered batch into two halves. Do not include the median into the halves.
  2. Q_1 is the median of the lower half of the data. Q_3 is the median of the upper half of the data.

- **MAD** is the median of the numbers $|x_i - \tilde{x}|$.

Example

Let the ordered batch be $\{x_i\} = \{1, 2, 5, 6, 9, 11, 19\}$

- $Q_2 = \tilde{x} = 6$
- $Q_1 = 2$
- $Q_3 = 11$

$$\mathrm{IQR} = 9$$

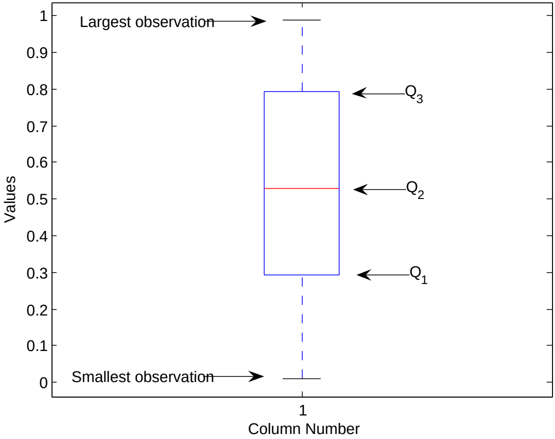
- $\{|x_i - \tilde{x}|\} = \{5, 4, 1, 0, 3, 5, 13\}$

$$\mathrm{MAD} = 4$$

Boxplots

A boxplot is a graphical display of numerical data that is based on five-number summaries: the smallest observation, lower quartile (Q_1), median (Q_2), upper quartile (Q_3), and largest observation.

Example: $x_1, \ldots, x_n \sim U[0,1]$, $n = 100$



Summary

- Measures of Location
  - Arithmetic Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers)
  - Median: the middle value of the ordered batch values $\tilde{x} = Q_2$
  - Trimmed Mean:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

  - M estimate: $y^* = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$
    - ★ if $\Psi(x_i, y) = (x_i - y)^2$, then $y^* = \overline{x}$
    - ★ it $\Psi(x_i, y) = |x_i - y|$, then $y^* = \tilde{x}$
- Measures of Dispersion
  - Sample Standard Deviation (sensitive to outliers):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  - Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$
  - Median Absolute Deviation: $\mathrm{MAD} =$ median of the numbers $|x_i - \tilde{x}|$
- Boxplots are useful graphical displays.

Math 408 - Mathematical Statistics

Lecture 38. Fundamental Concepts of Statistical Inference: an Overview

May 3, 2013

Agenda

- Probability Theory
- Survey Sampling
- Fundamental Concepts of Statistical Inference

Statistical Inference

Statistical inference is the process of using data to infer the distribution that generates the data. The basic statistical inference problem is the following:

Basic Problem

We observe $X_1, \ldots, X_n \sim \pi$. We want to estimate π or some features of π such as its mean.

Definition

A **statistical model** is a set of distributions or a set of densities $\mathcal{F}$.

- A **parametric model** is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
- A **nonparametric model** is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.

Point Estimation, Confidence Intervals, Hypothesis Testing

Given a parametric model, $\mathcal{F} = \{\pi(x|\theta), \ \ \theta \in \Theta\}$, the problem of inference is then to estimate the parameter θ from the data.

Almost all problems in statistical inference can be identified as being one of three types: **point estimates**, **confidence intervals**, and **hypothesis testing**.

- Point Estimation refers to providing a single "best guess."
  Suppose $X_1, \ldots, X_n \sim \pi(x|\theta)$, where $\pi(x|\theta) \in \mathcal{F}$.
  A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = f(X_1, \ldots, X_n)$$

- A $100(1-\alpha)\%$ Confidence Interval for a parameter θ is a random interval $I_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that

$$\mathbb{P}(\theta \in I_n) = 1 - \alpha$$

- In Hypothesis Testing, we start with some default theory, called a null hypothesis, and we ask if the data provide sufficient evidence to reject the theory. If not, we accept the null hypothesis.

Method of Moments

Suppose that $X_1, \ldots, X_n \sim \pi(x|\theta)$ where $\theta \in \Theta$, and we want to estimate θ based on the data $X_1, \ldots, X_n$.

Method of Moments

- *Let $\mu_j(\theta) = \mathbb{E}_\theta[X^j]$ be the j^{th} moment of a probability distribution $\pi(x|\theta)$*
- *Let $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ be the j^{th} sample moment*
  (LLN: $\hat{\mu}_j \xrightarrow{\mathbb{P}} \mu_j(\theta)$, when $n \to \infty$)
- *Suppose that the parameter θ has k components, $\theta = (\theta_1, \ldots, \theta_k)$*

The **method of moments estimator** *$\hat{\theta}$ is defined to be the value of θ such that*

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \quad \ldots\ldots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases} \tag{1}$$

- System (1) is a system of k equations with k unknowns: $\theta_1, \ldots, \theta_k$
- The solution of this system $\hat{\theta}$ is the MoM estimate of the parameter θ.

Consistency of the MoM estimator

Definition

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is **consistent** if

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

Theorem

The method of moments estimate is consistent.

The Likelihood Function

The most common method for estimating parameters in a parametric model is the method of maximum likelihood.

Suppose $X_1, \dots, X_n$ are i.i.d. from $\pi(x|\theta)$.

Definition

The **likelihood function** is defined by

$$\boxed{\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi(X_i|\theta)}$$

Important Remark:

- The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter θ.

Maximum Likelihood Estimate

Definition

The **maximum likelihood estimate** (MLE) of θ, denoted $\hat{\theta}_{\mathrm{MLE}}$, is the value of θ that maximizes the likelihood $\mathcal{L}(\theta)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

$\hat{\theta}_{\mathrm{MLE}}$ makes the observed data $X_1, \ldots, X_n$ "most probable" or "most likely"

Important Remark:

Rather than maximizing the likelihood itself, it is often easier to maximize its natural logarithm (which is equivalent since the log is a monotonic function). The log-likelihood is

$$\boxed{l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \pi(X_i|\theta)}$$

Properties of MLE

- MLE is consistent:
$$\hat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathbb{P}} \theta_0$$
where θ_0 denotes the true value of θ.
- MLE is equivariant:
if $\hat{\theta}_{\mathrm{MLE}}$ is the MLE of $\theta \quad \Rightarrow \quad f(\hat{\theta}_{\mathrm{MLE}})$ is the MLE of $f(\theta)$.
- MLE is asymptotically optimal: among all well behaved estimators, the MLE has the smallest variance, at least for large sample sizes n.
- MLE is asymptotically Normal:
$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$
where
$$I(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\pi(X|\theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log\pi(x|\theta)\right)^2\pi(x|\theta)dx$$
  - $I(\theta)$ is called Fisher Information.
- MLE is asymptotically unbiased:
$$\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_{\mathrm{MLE}}] = \theta_0$$

Confidence Intervals from MLEs

Recall that

Definition

A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a random interval calculated from the sample,

$$X_1, \ldots, X_n \sim \pi(x|\theta)$$

which contains θ with probability $1-\alpha$.

There are three methods for constructing confidence intervals using MLEs $\hat{\theta}_{\mathrm{MLE}}$:

- Exact Method
- Approximate Method
- Bootstrap Method

Exact Method

Exact Method provides exact confidence intervals.

- Example: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu : \quad \hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}} \hat{\sigma}^2_{\mathrm{MLE}} t_{n-1}(\alpha/2)$$

$$\sigma^2 : \quad \left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(1-\frac{\alpha}{2})} \right)$$

These result is based of the following facts:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Remark:

The main drawback of the exact method is that in practice the sampling distributions — like t_{n-1} and χ^2_{n-1} in our example — are not known.

Approximate Method

One of the most important properties of MLE is that it is asymptotically normal:

$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right), \qquad \text{as } n \to \infty$$

where $I(\theta_0)$ is Fisher information

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \pi(X|\theta) \right)^2 \right]$$

Since the true value θ_0 is unknown, we will use $I(\hat{\theta}_{\mathrm{MLE}})$ instead of $I(\theta_0)$:

Result

An **approximate** $100(1-\alpha)\%$ *confidence interval for* θ_0 *is*

$$\hat{\theta}_{\mathrm{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{\mathrm{MLE}})}}$$

where z_α *is the point beyond which the standard normal distribution has probability* α.

Measure of Efficiency: Mean Squared Error

In most estimation problems, there are many possible estimates $\hat{\theta}$ of θ. For example, the MoM estimate $\hat{\theta}_{\mathrm{MoM}}$ or the MLE estimate $\hat{\theta}_{\mathrm{MLE}}$.

Question: How would we choose which estimate to use?

Qualitatively, it is reasonable to choose that estimate whose distribution is most highly concentrated about the true parameter value θ_0. To make this idea work, we need to define a quantitative measure of such concentration.

Definition

The **mean squared error** of $\hat{\theta}$ as an estimate of θ_0 is

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]}$$

- The mean squared error can be also written as follows:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] + \underbrace{(\mathbb{E}(\hat{\theta}) - \theta_0)^2}_{\text{squared bias}}$$

- If $\hat{\theta}$ is unbiased, then $\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}]$.

Cramer-Rao Inequality

Let $X_1, \dots, X_n$ be i.i.d. from $\pi(x|\theta)$. Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be any unbiased estimate of a parameter θ whose true values is θ_0. Then, under smoothness assumptions on $\pi(x|\theta)$,

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}$$

Important Remarks:

- $\hat{\theta}$ can't have arbitrary small MSE
- The Cramer-Rao inequality gives a lower bound on the variance of any unbiased estimate.

Definition

An unbiased estimate whose variance achieves this lower bound is said to be **efficient**.

Recall that MLE is asymptotically Normal: $\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$

- Therefore, MLE is asymptotically efficient
- However, for a finite sample size n, MLE may not be efficient

Hypothesis Testing: General Framework

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**.

Let X be data and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $\mathcal{R} \subset \mathcal{X}$ called the **rejection region**. If $X \in \mathcal{R}$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$X \in \mathcal{R} \Rightarrow \text{ reject } H_0$$

$$X \notin \mathcal{R} \Rightarrow \text{ accept } H_0$$

Usually the rejection region $\mathcal{R}$ is of the form

$$\mathcal{R} = \{x \in \mathcal{X} : T(x) < c\}$$

where T is a **test statistic** and c is a **critical value**.
The main problem in hypothesis testing is

to find an appropriate test statistic T and an appropriate cutoff value c

Main Definitions

In hypothesis testing, there are two types of errors we can make:

- Rejecting H_0 when H_0 is true is called a **type I error**
- Accepting H_0 when H_1 is true is called a **type II error**

Definition

- The probability of a type I error is called the **significance level** of the test and is denoted by α

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(\text{Reject } H_0 | H_0)$$

- The probability of a type II error is is denoted by β

$$\beta = \mathbb{P}(\text{type II error}) = \mathbb{P}(\text{Accept } H_0 | H_1)$$

- $(1 - \beta)$ is called the **power** of the test

$$\textbf{power} = 1 - \beta = 1 - \mathbb{P}(\text{Accept } H_0 | H_1) = \mathbb{P}(\text{Reject } H_0 | H_1)$$

Thus, the **power** of the test is the probability of rejecting H_0 when it is false.

Neyman-Pearson Lemma

Definition

- A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**.
- A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**.

The Neyman-Pearson Lemma shows that the test that is based on the likelihood ratio is optimal for simple hypotheses:

Neyman-Pearson Lemma

Suppose that H_0 and H_1 are simple hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Suppose that the **likelihood ratio test** *that rejects H_0 whenever the likelihood ratio is less than c,*

$$\textit{Reject } H_0 \quad \Leftrightarrow \quad \frac{\mathcal{L}(Data|\theta_0)}{\mathcal{L}(Data|\theta_1)} < c$$

has significance level α_{LR}. Then any other test for which the significance level $\alpha \leq \alpha_{\mathrm{LR}}$ has power less than or equal to that of the likelihood ratio test

$$1 - \beta \leq 1 - \beta_{LR}$$

Generalized Likelihood Ratio Test

Let $X = (X_1, \ldots, X_n)$ be data and let $\pi(x|\theta)$ be the joint density of the data. The likelihood function is then

$$\mathcal{L}(\theta) = \pi(X|\theta)$$

Suppose we we wish to test

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

where Θ_0 and Θ_1 are two disjoint sets of the parameter space Θ, $\Theta = \Theta_0 \sqcup \Theta_1$.

- Based on the data, a measure of relative plausibility of the hypotheses is the ratio of their likelihoods.
- If the hypotheses are composite, each likelihood is evaluated at that value of θ that maximizes it.

This yields the **generalized likelihood ratio**:

$$\boxed{\Lambda^* = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta_1} \mathcal{L}(\theta)}}$$

Small values of Λ^* tend to discredit H_0.

Generalized Likelihood Ratio Test

For technical reasons, it is preferable to use the following statistic instead of Λ^*:

$$\boxed{\Lambda = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta} \mathcal{L}(\theta)}}$$

- Λ is called the likelihood ratio statistic.
- Note that

  $$\Lambda = \min\{\Lambda^*, 1\}$$

  Thus, small values of Λ^* correspond to small values of Λ.

The rejection region $\mathcal{R}$ for a generalized likelihood test has the following form:

$$\text{reject } H_0 \quad \Leftrightarrow \quad X \in \mathcal{R} = \{X : \Lambda(X) < \lambda\}$$

The threshold λ is chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha,$$

where α is the desired significance level of the test.

Distribution of $\Lambda(X)$

In order for the generalized likelihood ratio test to have the significance level α, the threshold λ must be chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha$$

If the distribution of $\Lambda(X)$ under H_0 is known, then we can determine λ.
Generally, the distribution of Λ is not of a simple form, but in many situations the following theorem provides the basis for an approximation of the distribution.

Theorem

Under smoothness conditions on $\pi(x|\theta)$, the null distribution of $-2\log\Lambda(X)$ (i.e. distribution under H_0) tends to a χ^2_d as the sample size $n \to \infty$, where

$$d = \dim\Theta - \dim\Theta_0,$$

where $\dim\Theta$ *and* $\dim\Theta_0$ *are the numbers of free parameters in Θ and Θ_0.*

Summarizing Data: Empirical CDF

Suppose that $x_1, \dots, x_n$ is a batch of numbers.

Remark: We use the word

- "sample" when $X_1, \dots, X_n$ is a collection of random variables.
- "batch" when $x_1, \dots, x_n$ are fixed numbers (data, realization of sample).

Definition

The **empirical cumulative distribution function** (eCDF) is defined as

$$\boxed{F_n(x) = \frac{1}{n}(\# x_i \leq x)}$$

Denote the ordered batch of numbers by $x_{(1)}, \dots, x_{(n)}$.

- If $x < x_{(1)}$, then $F_n(x) = 0$
- If $x_{(1)} \leq x < x_{(2)}$, then $F_n(x) = 1/n$
- If $x_{(k)} \leq x < x_{(k+1)}$, then $F_n(x) = k/n$

The eCDF is the "data analogue" of the CDF of a random variable

Summarizing Data: Quantile-Quantile Plots

Quantile-Quantile (Q-Q) plots are used for comparing two probability distributions.

Suppose that X is a continuous random variable with a strictly increasing CDF F.

Definition

The p^{th} **quantile** of F is that value x_p such that

$$F(x_p) = p \quad \text{or} \quad \boxed{x_p = F^{-1}(p)}$$

Suppose we want to compare two CDF: F and G.

Definition

The **theoretical Q-Q plot** is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$, that is the graph $[F^{-1}(p), G^{-1}(p)]$ for $p \in (0, 1)$.

- If the two CDFs are identical, the theoretical Q-Q plot will be the line $y = x$.

Summarizing Data: Empirical Q-Q plots

In practice, a typical scenario is the following:

- $F(x) = F_0(x)$ is a specified CDF (e.g. normal) which is a theoretical model for data $X_1, \ldots, X_n$.
- $G(x)$ is the empirical CDF for $x_1, \ldots, x_n$, a realization of $X_1, \ldots, X_n$ (actually observed data).
- We want to compare the model $F(x)$ with the observation $G(x)$.

Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered batch. Then

Definition

The **empirical Q-Q plot** is the plot of $F_0^{-1}(i/n)$ on the horizonal axis versus $G^{-1}(i/n) = x_{(i)}$ on the vertical axis, for $i = 1, \ldots, n$.

Remarks:

- The quantities $p_i = i/n$ are called plotting positions

Summarizing Data: Measures of Location and Dispersion

- Measures of Location
  - Arithmetic Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers)
  - Median: the middle value of the ordered batch values $\tilde{x} = Q_2$
  - Trimmed Mean:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

  - M estimate: $y^* = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$
    - ★ if $\Psi(x_i, y) = (x_i - y)^2$, then $y^* = \overline{x}$
    - ★ it $\Psi(x_i, y) = |x_i - y|$, then $y^* = \tilde{x}$
- Measures of Dispersion
  - Sample Standard Deviation (sensitive to outliers):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  - Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$
  - Median Absolute Deviation: $\mathrm{MAD} =$ median of the numbers $|x_i - \tilde{x}|$

Thank you for attention and good luck on the final!