Introduction to Statistics

CS 3130 / ECE 3530: Probability and Statistics for Engineers

March 14, 2024

Independent, Identically Distributed RVs

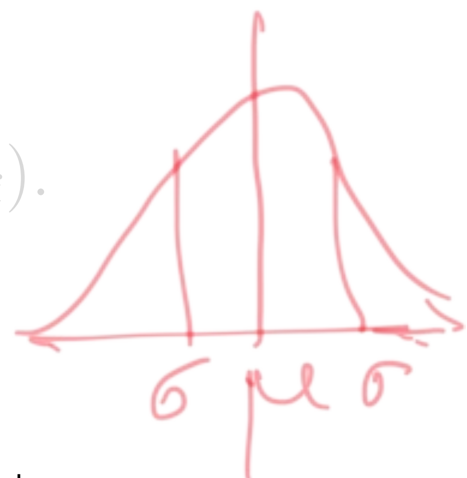
Definition

The random variables $X_1, X_2, \ldots, X_n$ are said to be **independent, identically distributed (iid)** if they share the same probability distribution and are independent of each other.

Independence of n random variables means

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

$X_1 \; \ldots \; X_n = \prod f_{X_i}(\;)$



Independent, Identically Distributed RVs

Definition

The random variables $X_1, X_2, \ldots, X_n$ are said to be **independent, identically distributed (iid)** if they share the same probability distribution and are independent of each other.

Independence of n random variables means

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

Random Samples

$\underbrace{X_1, \ldots, X_n}_{\text{Random samples}}$

Definition

A **random sample** from the distribution F of length n is a set $(X_1, \ldots, X_n)$ of iid random variables with distribution F. The length n is called the **sample size**.

- A random sample represents an experiment where n independent measurements are taken.
- A **realization** of a random sample, denoted $(x_1, \ldots, x_n)$ are the values we get when we take the measurements.

Random Samples

Definition

A **random sample** from the distribution F of length n is a set $(X_1, \ldots, X_n)$ of iid random variables with distribution F. The length n is called the **sample size**.

- A random sample represents an experiment where n independent measurements are taken.
- A **realization** of a random sample, denoted $(x_1, \ldots, x_n)$ are the values we get when we take the measurements.

Random Samples

Definition

A **random sample** from the distribution F of length n is a set $(X_1, \ldots, X_n)$ of iid random variables with distribution F. The length n is called the **sample size**.

- A random sample represents an experiment where n independent measurements are taken.
- A **realization** of a random sample, denoted $(x_1, \ldots, x_n)$ are the values we get when we take the measurements.

$x_1, \ldots, x_n$

Statistics

Definition

A **statistic** on a random sample $(X_1, \ldots, X_n)$ is a function $T(X_1, \ldots, X_n)$.

Examples:

- Sample Mean

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

Statistics

Definition

A **statistic** on a random sample $(X_1, \ldots, X_n)$ is a function $T(X_1, \ldots, X_n)$.

Examples:

- Sample Mean

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

$X \sim P$

$$E(X) = \sum_{i=1}^{n} x_i \, P(x_i) = \sum x_i \, f_X(x_i)$$

$$= \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum x_i$$

$$= \frac{1}{n} \sum X_i$$

Statistics

Definition

A **statistic** on a random sample $(X_1, \ldots, X_n)$ is a function $T(X_1, \ldots, X_n)$.

Examples:

- Sample Mean

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

- Sample Variance

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

$$\underline{\mathrm{Var}(X) = E\left[(X - E(X))^2\right]}$$

$$E(X) = \overline{X}_m \qquad X \sim p$$

$$= E\left[(X - \overline{X}_m)^2\right]$$

$$= \sum_{i=1}^{m} (X_i - \overline{X}_m)\, \underline{p(X_i)}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (X_i - \overline{X}_m)^2$$

$$x_1, x_2, x_3, x_4$$

$$5 \quad 6 \quad 6 \quad 7$$

$$\bar{x}_m = \frac{1}{4}(5+6+6+7)$$

$$\bar{x}_m = \frac{24}{4} = 6$$

$$s_n^2 = \frac{1}{3}\left[(5-6)^2 + \cancel{(6-6)^2} + \cancel{(6-6)^2} + (7-6)^2\right]$$

$$= \frac{1}{3}[1+1] = \frac{2}{3} = 0.67$$

Order Statistics

50, 30, 10, 40, 20
X_1 . . . X_5
10 20 30 40 50

Given a sample $X_1, X_2, \ldots, X_n$, start by sorting the list of numbers.

- The **median** is the center element in the list if n is odd, average of two middle elements if n is even.
- The i**th order statistic** is the ith element in the list.
- The **empirical quantile** $q_n(p)$ is the first point at which p proportion of the data is below.
- **Quartiles** are $q_n(p)$ for $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$. The **inner-quartile range** is $IQR = q_n(0.75) - q_n(0.25)$.

Order Statistics

10 20 30 40 50

Given a sample $X_1, X_2, \ldots, X_n$, start by sorting the list of numbers.

- The **median** is the center element in the list if n is odd, average of two middle elements if n is even.
- The i**th order statistic** is the ith element in the list.
- The **empirical quantile** $q_n(p)$ is the first point at which p proportion of the data is below.
- **Quartiles** are $q_n(p)$ for $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$. The **inner-quartile range** is $IQR = q_n(0.75) - q_n(0.25)$.

Order Statistics

10 20 30 40 50
1/5 2/5 3/5 4/5 5/5
0.2 0.4 0.6 0.8 1.0

Given a sample $X_1, X_2, \ldots, X_n$, start by sorting the list of numbers.

0.3

- The **median** is the center element in the list if n is odd, average of two middle elements if n is even.
- The **ith order statistic** is the ith element in the list.

$p = 0.75$

- The **empirical quantile** $q_n(p)$ is the first point at which p proportion of the data is below.
- **Quartiles** are $q_n(p)$ for $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$. The **inner-quartile range** is $IQR = q_n(0.75) - q_n(0.25)$.

Order Statistics

10, 20, 30, 40, 50
0.25 0.5 0.75 1

Given a sample $X_1, X_2, \ldots, X_n$, start by sorting the list of numbers.

- The **median** is the center element in the list if n is odd, average of two middle elements if n is even.
- The i**th order statistic** is the ith element in the list.
- The **empirical quantile** $q_n(p)$ is the first point at which p proportion of the data is below.
- **Quartiles** are $q_n(p)$ for $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$. The **inner-quartile range** is
  $IQR = q_n(0.75) - q_n(0.25)$.

0.25, 0.5, 0.75, 1

$40 - 20 = 20$

Realizations of Statistics

$\bar{x}_n$

$2{:}5 \quad X_1, X_2, X_3, X_4$

$\bar{x}_n' = (1 \quad 2 \quad 3 \quad 4)$

$\bar{x}_n'' = (3 \quad 10 \quad 7 \quad 5)$

Remember, a statistic is a random variable! It is not a fixed number, and it has a distribution.

$\frac{25}{4} = 6.25$ ✓

If we perform an experiment, we get a realization of our sample $(x_1, x_2, \ldots, x_n)$. Plugging these numbers into the formula for our statistic gives a **realization of the statistic**, $t = T(x_1, x_2, \ldots, x_n)$.

Example: given realizations x_i of a random sample, the realization of the sample mean is $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Upper-case = random variable, Lower-case = realization

Realizations of Statistics

Remember, a statistic is a random variable! It is not a fixed number, and it has a distribution.

If we perform an experiment, we get a realization of our sample $(x_1, x_2, \ldots, x_n)$. Plugging these numbers into the formula for our statistic gives a **realization of the statistic**, $t = T(x_1, x_2, \ldots, x_n)$.

Example: given realizations x_i of a random sample, the realization of the sample mean is $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Upper-case = random variable, Lower-case = realization

Realizations of Statistics

Remember, a statistic is a random variable! It is not a fixed number, and it has a distribution.

If we perform an experiment, we get a realization of our sample $(x_1, x_2, \ldots, x_n)$. Plugging these numbers into the formula for our statistic gives a **realization of the statistic**, $t = T(x_1, x_2, \ldots, x_n)$.

Example: given realizations x_i of a random sample, the realization of the sample mean is $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

X x

Upper-case = random variable, Lower-case = realization

Realizations of Statistics

Remember, a statistic is a random variable! It is not a fixed number, and it has a distribution.

If we perform an experiment, we get a realization of our sample $(x_1, x_2, \ldots, x_n)$. Plugging these numbers into the formula for our statistic gives a **realization of the statistic**, $t = T(x_1, x_2, \ldots, x_n)$.

Example: given realizations x_i of a random sample, the realization of the sample mean is $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Upper-case = random variable, Lower-case = realization

Statistical Plots

(See example code “StatPlots.r”)

- Histograms
- Empirical CDF
- Box plots
- Scatter plots

Sampling Distributions

Given a sample $(X_1, X_2, \ldots, X_n)$. Each X_i is a random variable, all with the same pdf. identical

And a statistic $T = T(X_1, X_2, \ldots, X_n)$ is also a random variable and has its own pdf (different from the X_i pdf). This distribution is the **sampling distribution** of T.

If we know the distribution of the statistic T, we can answer questions such as "What is the probability that T is in some range?" This is $P(a \leq T \leq b)$ – computed using the cdf of T.

Sampling Distributions

Given a sample $(X_1, X_2, \ldots, X_n)$. Each X_i is a random variable, all with the same pdf.

And a statistic $T = T(X_1, X_2, \ldots, X_n)$ is also a random variable and has its own pdf (different from the X_i pdf). This distribution is the **sampling distribution** of T.

If we know the distribution of the statistic T, we can answer questions such as "What is the probability that T is in some range?" This is $P(a \leq T \leq b)$ – computed using the cdf of T.

Sampling Distributions

Given a sample $(X_1, X_2, \ldots, X_n)$. Each X_i is a random variable, all with the same pdf.

And a statistic $T = T(X_1, X_2, \ldots, X_n)$ is also a random variable and has its own pdf (different from the X_i pdf). This distribution is the **sampling distribution** of T.

If we know the distribution of the statistic T, we can answer questions such as “What is the probability that T is in some range?” This is $\boxed{P(a \leq T \leq b)}$ – computed using the cdf of T.

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

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

- It's expectation is $E[\bar{X}_n] = \mu$
- It's variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

$\sigma^2 \quad \sigma^2 \quad \sigma^2$

$X_1, X_2, \ldots, X_n$

$\mu \quad \mu \quad \mu$

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

- Its expectation is $E[\bar{X}_n] = \mu$.
- Its variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

$\bar{x}_n^1$ $\quad$ $\bar{x}_n^2$ $\quad$ $\bar{x}_n^3$

$x_1 \ldots x_n$ $\quad$ $x_1 \ldots x_n$ $\quad$ $x_1 \ldots x_n$

$\langle x_1 \ldots x_n \rangle$ $\quad$ $\langle \quad \rangle$ $\quad$ $\langle \quad \rangle$

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} E\left[\left(\sum_{i=1}^{n} X_i\right)\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{\cancel{n}}\cancel{n}\mu = \mu$$

- It's expectation is $E[\bar{X}_n] = \mu$
- It's variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

- It's expectation is $E[\bar{X}_n] = \mu$
- It's variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1\cancel{n}\sigma^2}{n^{\cancel{2}}} = \frac{\sigma^2}{n}$$

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

- It's expectation is $E[\bar{X}_n] = \mu$
- It's variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

Sampling Distribution of the Mean

Given a sample $(X_1, X_2, \ldots, X_n)$ with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

What do we know about the distribution of the sample mean, $\bar{X}_n$?

- It's expectation is $E[\bar{X}_n] = \mu$
- It's variance is $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- As n get's large, it is approximately a Normal distribution with mean μ and variance σ^2/n.
- Not much else! We don't know the full pdf/cdf.

2

$x_1 \to 1$

$x_2 \to 2$

$x_3 \to 3$

$x_4 \to 4$

	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)
4	(4,1)	(4,2)	(4,3)	(4,4)

2	3	4	5
3	4	5	6
4	5	6	7
5	6	7	8

$2 \to 1/16$

$3 \to 2/16$

$4 \to 3/16$

$5 \to 4/16$

$6 \to 3/16$

$7 \to 2/16$

$8 \to 1/16$

When the X_i are Normal

μ, σ

When the sample is Normal, i.e., $X_i \sim N(\mu, \sigma^2)$, then we know the *exact* sampling distribution of the mean $\bar{X}_n$ is Nornal:

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

$N(\mu, \sigma^2/n)$

Chi-Square Distribution

$X \sim N(0,1)$

The **chi-square distribution** is the distribution of a sum of squared Normal random variables. So, if $X_i \sim N(0,1)$ are iid, then

X_1 $x_1 = 0.25$

X_2 $x_2 = 0.9$

$x_3 = -0.75$

$$Y = \sum_{i=1}^{k} X_i^2$$

has a chi-square distribution with k **degrees of freedom**. We write $Y \sim \chi^2(k)$.

Read the Wikipedia page for this distribution!!