Introduction to Statistics

CS 3130 / ECE 3530: Probability and Statistics for Engineers

March 21, 2023
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

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

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.
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**
  \[
  S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
  \]
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**

  \[
  S^2_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
  \]
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**
  \[
  S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
  \]
Order Statistics

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 $i$th 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 $IQR = q_n(0.75) - q_n(0.25)$. 

University of Utah, CS3130, Spring 2024, Prof. Ross Whitaker
Order Statistics

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 $i$th 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

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 $i$th 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

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 $i$th 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 $IQR = q_n(0.75) - q_n(0.25)$. 
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 \(\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i\).

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 \(\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i\).
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.

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 \(P(a \leq T \leq b)\) – 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 $\text{Var}(X_i) = \sigma^2$,

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 $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
▶ As $n$ gets large, it is approximately a Normal distribution with mean $\mu$ and variance $\sigma^2/n$.
▶ 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 \(\text{Var}(X_i) = \sigma^2\),

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 \(\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\)
- As \(n\) gets large, it is approximately a Normal distribution with mean \(\mu\) and variance \(\sigma^2/n\).
- 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 \(\text{Var}(X_i) = \sigma^2\),

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 \(\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\)
- As \(n\) gets large, it is approximately a Normal distribution with mean \(\mu\) and variance \(\sigma^2/n\).
- 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 \(\text{Var}(X_i) = \sigma^2\),

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 \(\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\)
- As \(n\) get’s large, it is approximately a Normal distribution with mean \(\mu\) and variance \(\sigma^2/n\).
- 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 \(\text{Var}(X_i) = \sigma^2\),

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 \(\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\)
- As \(n\) get’s large, it is approximately a Normal distribution with mean \(\mu\) and variance \(\sigma^2/n\).

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 \(\text{Var}(X_i) = \sigma^2\),

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 \(\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\)
- As \(n\) gets large, it is approximately a Normal distribution with mean \(\mu\) and variance \(\sigma^2/n\).
- Not much else! We don’t know the full pdf/cdf.
When the $X_i$ are Normal

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 Normal:

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

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

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

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

Read the Wikipedia page for this distribution!!
Chi-Square Distribution

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

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

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

Read the Wikipedia page for this distribution!!
Sampling Distribution of the Variance

If $X_i \sim N(\mu, \sigma)$ are iid Normal RV's, then the sample variance is distributed as a scaled chi-square random variable:

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

Or, a slight abuse of notation, we can write:

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

This means that the $S_n^2$ is a chi-square random variable that has been scaled by the factor $\frac{\sigma^2}{n-1}$.
How to Scale a Random Variable

Let’s say I have a random variable $X$ that has pdf $f_X(x)$.

What is the pdf of $kX$, where $k$ is some scaling constant?

The answer is that $kX$ has pdf

$$f_{kX}(x) = \frac{1}{k} f_X\left(\frac{x}{k}\right)$$

See pg 106 (Ch 8) in the book for more details.
Central Limit Theorem

Let $X_1, X_2, \ldots$ be iid random variables from a distribution with mean $\mu$ and variance $\sigma^2 < \infty$. Then in the limit as $n \to \infty$, the statistic

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

has a standard normal distribution.

Recall $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. 
Importance of the Central Limit Theorem

- Applies to real-world data when the measured quantity comes from the average of many small effects.
- Examples include electronic noise, interaction of molecules, exam grades, etc.
- This is why a Normal distribution model is often used for real-world data.
- Also, this “concentration of measure” effect is the basis for all of machine learning (more data, more accuracy).
Importance of the Central Limit Theorem

▶ Applies to real-world data when the measured quantity comes from the average of many small effects.
▶ Examples include electronic noise, interaction of molecules, exam grades, etc.
▶ This is why a Normal distribution model is often used for real-world data.
▶ Also, this “concentration of measure” effect is the basis for all of machine learning (more data, more accuracy).
Importance of the Central Limit Theorem

- Applies to real-world data when the measured quantity comes from the average of many small effects.
- Examples include electronic noise, interaction of molecules, exam grades, etc.
- This is why a Normal distribution model is often used for real-world data.
- Also, this “concentration of measure” effect is the basis for all of machine learning (more data, more accuracy).
Importance of the Central Limit Theorem

- Applies to real-world data when the measured quantity comes from the average of many small effects.
- Examples include electronic noise, interaction of molecules, exam grades, etc.
- This is why a Normal distribution model is often used for real-world data.
- Also, this “concentration of measure” effect is the basis for all of machine learning (more data, more accuracy).