Facultad de Ciencias Exactas | UNLP

Facultad de INFORMÁTICA | UNIVERSIDAD NACIONAL DE LA PLATA

Especialización en Bioinformática

Manejo de datos en biología computacional. Herramientas de Estadística

Año 2022

Duración total: 70hs.

Responsable: Facultad de Ciencias Exactas

Docente Responsable: Dr. Leandro Sommese

OBJETIVO GENERAL

El objetivo general del curso es sistematizar y aplicar el diseño de experimentos, el manejo automatizado de datos y el análisis estadístico utilizando lenguajes de programación.

Objetivos específicos:

- Favorecer el uso de herramientas de programación que faciliten el trabajo de investigación en ciencias biológicas a través de la automatización de procesos.
- Incorporar conceptos generales de programación para descubrir y comprender el potencial de la actividad.
- Introducir conceptos propios de herramientas computacionales relevantes para el estudio de sistemas biológicos.
- Introducir el uso de funciones específicas para el tratamiento de datos biológicos y para la presentación de datos para publicaciones científicas.
- Fomentar la práctica contextualizada en ejemplos biológicos concretos, de forma intensa y constante, como vehículo esencial para incorporar la programación como herramienta útil.
- Problematizar sobre la necesidad y utilidad de la Estadística como herramienta en su ejercicio profesional.
- Introducir las diferentes reglas de las probabilidades para la toma de decisiones y modelos de distribuciones probabilísticas.
- Aplicar adecuadamente las técnicas de la Estadística Inferencial.
- Seleccionar el estadístico más apropiado para cada tipo de inferencia inductiva.
- Calcular el tamaño adecuado de muestra a seleccionar, con cierto nivel de riesgo o significancia.
- Aplicar las técnicas de Test de Hipótesis, sobre los parámetros poblacionales, para una y dos poblaciones.

- Aplicar apropiadamente las Pruebas de Estadística no Paramétrica, para poblaciones, relacionadas o independientes y reconocer que Test de Hipótesis no paramétrico es el indicado para testear distribuciones libres en escalas nominal, ordinal e interválica.

COMPETENCIAS A DESARROLLAR EN RELACION CON EL OBJETIVO DE LA CARRERA

C. 2 - Utilizar distintas técnicas de procesamiento de datos biológicos para su representación/visualización y análisis eficiente, mediante algoritmos de software ejecutados sobre plataformas adecuadas para el tipo y volumen de datos en cuestión.

PROGRAMA

Módulo I: Introducción a lenguajes utilizados para el manejo de datos

- Generalidades. Sintaxis básica. Tipos de objetos. Definición y uso de variables. Operadores. Estructuras de control. Definición y uso de funciones. Uso de bibliotecas.
- Manejo de tablas de datos. Lectura y escritura de datos estructurados. Agrupamiento de datos. Manejo de datos faltantes. Distribución de frecuencias. Medidas de distribución (centralización y dispersión).

Módulo II: Análisis de datos

- Estadística descriptiva: Población y muestra. Caracteres cuantitativos o cualitativos. Variable estadística. Distribuciones de frecuencias. Tabla de frecuencias de una variable discreta. Agrupamiento en intervalos de clase.
- Medidas características de una distribución: Medidas de centralización. Medidas de dispersión.
- Recorridos. Desviación media. Varianza y desviación típica. Coeficientes de variación. Asimetría y curtosis.
- Variables aleatorias: Descripción de las variables aleatorias. Concepto de variable aleatoria. Variable aleatoria discreta. Variable aleatoria continua. Medidas características de una variable aleatoria. Varianza y desviación típica.
- Distribuciones discretas de probabilidad. Distribuciones continuas de probabilidad. Distribución χ de Pearson. Distribución t de Student. Distribución F de Fisher.

Módulo III: Inferencia estadística

- Muestreo. Media muestral. Varianza muestral. Estimación puntual de parámetros. Estimación por intervalos de confianza. Determinación del tamaño de la muestra.
- Módulo IV: Contraste de hipótesis
- Ensayos de hipótesis. Tipos de errores y significación. Contrastes bilaterales y unilaterales. Contraste de la media de una población normal. Contraste de una proporción. Contraste de varianza de una población normal. Contrastes de hipótesis para dos poblaciones. Aplicaciones de la distribución $\chi 2$. Análisis de varianza.
- Módulo V: Regresión lineal
- Regresión lineal simple. Ajuste de una recta de regresión. Correlación lineal. Interpretación del
- coeficiente de correlación. Inferencia estadística sobre la regresión.

- Módulo VI: Visualización de datos
- Generación de gráficos. Representaciones gráficas para datos agrupados. Representaciones gráficas para variables cualitativas. Comparación de bibliotecas. disponibles: matplotlib, seaborn, ggplot. Gráficos de dispersión, barra, líneas, caja y bigotes, histogramas, y mapas de color.
- Distribución y varianza muestral. Regresión lineal y ajuste a una recta. Coeficientes de regresión y correlación. Personalización de gráficos. Gráficos para publicación.

Módulo VII: Machine Learning - Clasificación, Regresión Lineal, Clustering

- Aprendizaje. Ciclo de trabajo de un Modelo Predictivo. Armado del Dataset. Entrenamiento y testeo. Overfitting y underfitting. Árboles de Decisión. Regresión. Logística. Curvas ROC. Evaluación de Resultados.
- Fundamentos de la Regresión Lineal. Supuestos y estimación de parámetros. Selección de Variables. Adecuación. Reducción de la dimensionalidad. Validación. Predicción.
- Definición de cluster. Aplicación en el campo de Datamining. Evaluación de similitudes. Medidas de Distancia. Clustering jerárquico. Evaluación de clusters. Validación de resultados.
- Módulo VIII: Modelos.
- Aplicación del análisis probabilístico e inferencial sobre modelos biológicos

METODOLOGIA Y MODALIDAD DE EVALUACION

El dictado de la asignatura corresponde a un curso teórico práctico de 6 horas semanales de cursada. Se trata de un curso teórico-práctico, en el cual, se dictan clases teóricas a cargo del docente, las que se intercalan con el desarrollo de los seminarios de manera tal que se conviertan en coloquios con el profesor y no en una mera exposición del mismo.
La aprobación de este curso requiere que las/los estudiantes entreguen y aprueben un trabajo práctico escritos para cada módulo. En caso de aprobación de los trabajos prácticos, la/el estudiante debe ofrecer un coloquio sobre un tema seleccionado y a partir de un diseño de un trabajo práctico final. La evaluación incluye un seguimiento constante.

ACTIVIDADES EXPERIMENTALES y DE INVESTIGACION

Las actividades a desarrollar son teorías y seminarios, en las cuales se desarrollan los contenidos del programa, intercalado con la explicación de los problemas fundamentales de los seminarios. Las clases teóricas no tienen un horario fijo, sino que se intercalan en el horario del teórico-práctico de acuerdo al transcurso del tema; las mismas se desarrollan de manera interactiva, de modo tal que los/las estudiantes tengan participación real a través de interrogantes planteados por el profesor. Cada eje central se desarrolla a partir de un trabajo práctico. Estos trabajos se evaluarán por su grado de concreción.

BIBLIOGRAFÍA BASICA

- Time Series Analysis: With Applications in R (2008). Cryer, Jonathan & Chan, Kung-Sik.

- Estadística Básica para Estudiantes de Ciencias (2009). J. Gorgas, N. Cardiel, J, Zamorano.

- Introduction to Statistics and Data Analysis (2016). C. Heumann, M. Schomaker, S. Shalabh.

- An Introduction to Machine Learning (2017). M. Kubat. Springer International Publishing, Second Edition.