

# **Advanced Biostatistics in Public Health (EP2004)**

Epidemiology & Medical Statistics Department

# Course Main Objective

- ✓ To collect, organize, summarize, analyze and present data.
- ✓ Determine the appropriate test to use based on how the data were collected and on the outcome variable of interest and be able to interpret results from that test.
- ✓ Use a statistical software tool (SPSS or R) to analyze data



# Students Assessment

#	Assessment task	Week Due	Percentage of Total Assessment Score
1	Final Examination	Week 12	50%
2	Practical Examination	Week 11	20%
3	Midterm Examination	Week 5	20%
4	Quiz	Week 3	10%
<b>Total</b>			<b>100%</b>





## What is Statistics?

- Statistics is the science of collecting, organizing, analyzing, interpreting and presenting data.

ما هي الإحصاءات؟

الإحصاءات هي علم جمع البيانات وتنظيمها وتحليلها وتفسيرها وتقديمها.

# Biostatistics

- The methods used in dealing with statistics in the fields of medicine, biology, and public health, and in planning, conducting and analyzing data which arise in investigations in these branches.

علم الإحصاء الحيوي (Biostatistics)

الإحصاء الحيوي هو الأساليب المستخدمة للتعامل مع الإحصاءات في مجالات الطب وعلم الأحياء والصحة العامة. ويشمل تخطيط، وتنفيذ، وتحليل البيانات الناتجة عن الأبحاث في هذه المجالات.



# Descriptive Statistics in Public Health

- **Definition**

- Methods for summarizing the data collected from a population or sample.

Examples:

**Mean:** Average number of cigarettes smoked per day by a group of individuals.

**Median:** The middle value in the number of doctor visits per year in a population.

**Mode:** The most common blood type in a group of patients.

الإحصاء الوصفي في الصحة العامة (Descriptive Statistics in Public Health)

التعريف:

هي الأساليب المستخدمة لتلخيص البيانات التي تم جمعها من مجتمع أو عينة.

أمثلة:

1. المتوسط (Mean):

عدد السجائر التي يدخنها مجموعة من الأفراد في اليوم.

2. الوسيط (Median):

القيمة الوسطى لعدد زيارات الطبيب في السنة داخل مجموعة سكانية.

3. النمط (Mode):

فصيلة الدم الأكثر شيوعًا في مجموعة من المرضى.

# Inferential Statistics in Public Health

الإحصاء الاستدلالي في الصحة العامة (Inferential Statistics in Public Health)

التعريف:

هي تقنيات تسمح لنا بتعميم النتائج على مجتمع كامل بناءً على بيانات مأخوذة من عينة.

أمثلة:

1. اختبار الفرضيات (Hypothesis Testing):

اختبار ما إذا كان الدواء الجديد يخفض ضغط الدم بشكل أكثر فعالية من العلاج الحالي باستخدام بيانات مأخوذة من عينة.

2. فواصل الثقة (Confidence Intervals):

تقدير النطاق الذي يقع فيه متوسط مستوى الكوليسترول الحقيقي في المجتمع استنادًا إلى بيانات عينة.

- **Definition**

Techniques that allow us to make generalizations about a population based on sample data.

Examples:

**Hypothesis Testing:** Testing whether a new drug reduces blood pressure more effectively than the current standard treatment based on sample data.

**Confidence Intervals:** Estimating the range in which the true average cholesterol level of a population lies, based on sample data.



# Sampling in Biostatistics

Key Methods, Data Collection, and  
Preparation for Analysis

العينة في الإحصاء الحيوي

المفاهيم الرئيسية: طرق جمع البيانات والاستعداد للتحليل



# Descriptive Statistics in Public Health

- **Definition**
  - Methods for summarizing the data collected from a population or sample.

الإحصاء الوصفي في الصحة العامة

التعريف:

طرق لتلخيص البيانات التي تم جمعها من مجتمع كامل أو عينة.

# Inferential Statistics in Public Health

- **Definition**

Techniques that allow us to make generalizations about a population based on sample data.

الإحصاء الاستدلالي في الصحة العامة

التعريف:

تقنيات تتيح لنا التعميم على المجتمع استنادًا إلى بيانات مأخوذة من عينة.

allow

# Population vs. Sample

## المجتمع مقابل العينة

- المجتمع (Population): المجموعة الكاملة التي ترغب في دراستها (مثل جميع البالغين في بلد معين).
- العينة (Sample): مجموعة صغيرة مختارة من المجتمع (مثل 1000 شخص تم استبيانهم حول عاداتهم في ممارسة الرياضة).

- **Population:** The entire group you want to study (e.g., all adults in a country).
- **Sample:** A smaller group selected from the population (e.g., 1,000 adults surveyed about their exercise habits).



# Definition of Sampling

تعريف أخذ العينات



Sampling: The process of selecting a subset of individuals from a population to represent the entire population.



Purpose: To draw conclusions about the population without studying everyone.

- أخذ العينات (Sampling): عملية اختيار مجموعة فرعية من الأفراد من المجتمع لتمثيله بالكامل.
- الهدف: استخلاص استنتاجات حول المجتمع دون الحاجة لدراسة الجميع.



# Importance of Sampling in Biostatistics

- يقلل الوقت والتكلفة.
  - يتيح دراسة المجتمعات الكبيرة.
  - يوفر تقديرات دقيقة وموثوقة.
  - يساهم في مراعاة الاعتبارات الأخلاقية في البحث.
- Reduces time and cost.
  - Allows for the study of large populations.
  - Provides accurate and reliable estimates.
  - Facilitates ethical considerations in research.

# Sampling Methods

## طرق أخذ العينات

1. العينة العشوائية البسيطة (Simple Random Sampling): لكل فرد فرصة متساوية في الاختيار.
2. العينة الطبقية (Stratified Sampling): يتم تقسيم المجتمع إلى طبقات، ويتم أخذ عينات من كل طبقة.
3. العينة العنقودية (Cluster Sampling): يتم اختيار مجموعات عشوائية من التجمعات الطبيعية.
4. العينة المنتظمة (Systematic Sampling): اتباع نظام معين في اختيار العينات (مثل اختيار كل عاشر شخص).

- **Simple Random Sampling:** Each individual has an equal chance of being selected.
- **Stratified Sampling:** Population is divided into strata by the researcher, and samples are drawn from each.
- **Cluster Sampling:** Random groups are selected from natural clusters.
- **Systematic Sampling:** Follow a specific system in sampling.

# Types of Data Collection

- **Printed Questionnaires:**

- Advantages: No need for internet, personal distribution possible.
- Disadvantages: Time-consuming, manual entry needed.

- **Electronic Questionnaires** (e.g., Microsoft Forms):

- Advantages: Automated data collection, easy to distribute.
- Disadvantages: Requires internet access, may exclude non-tech users.

## أنواع جمع البيانات

### 1. الاستبيانات الورقية:

- المزايا: لا حاجة للإنترنت، يمكن توزيعها شخصيًا.

- العيوب: تستغرق وقتًا طويلًا، وتتطلب إدخال البيانات يدويًا.

### 2. الاستبيانات الإلكترونية (مثل Microsoft Forms):

- المزايا: جمع البيانات تلقائيًا، وسهولة التوزيع.

- العيوب: تتطلب اتصالًا بالإنترنت، وقد تستثني المستخدمين غير المتمرسين تقنيًا.

البيانات المبلغ عنها ذاتيًا مقابل البيانات التي يجمعها المقابل

# Self-Reported vs. Interviewer-Collected Data

1. البيانات المبلغ عنها ذاتيًا (Self-Reported Data):

- يكمل المشاركون الاستبيانات بشكل مستقل.
- المخاطر: سوء فهم الأسئلة أو عدم الدقة.

2. البيانات التي يجمعها المقابل (Interviewer-Collected Data):

- يتم تنفيذ الاستبيان بواسطة مقاييلين مدربين.
- يقلل من سوء الفهم ولكنه قد يسبب تحيزًا بسبب تأثير المقابل.



## • Self-Reported Data:



- Participants complete questionnaires independently.



- Risk of bias due to misunderstanding questions or dishonesty.



## • Interviewer-Collected Data:



- Trained interviewers administer the questionnaire.



- Reduces misunderstanding but may introduce interviewer bias.




## إدخال البيانات في Excel

### 1. الإدخال اليدوي من الاستبيانات:

- إدخال الردود صفًا بصف لكل مشارك.
- تنظيم البيانات في أعمدة منظمة (مثل: رقم الهوية، العمر، النوع).

### 2. تصدير البيانات من الأنظمة الإلكترونية:

- تصدير البيانات مباشرة إلى Excel (مثل تنسيق CSV من Microsoft Forms).
- يوفر الوقت ويقلل من أخطاء النسخ.




## Entering Data into Excel

- **Manual Entry** from Questionnaires:
  - Enter responses row-by-row for each participant.
  - Organize data into structured columns (e.g., ID, age, gender).
- Exporting from **Electronic Platforms**:
  - Export data directly into Excel (e.g., CSV format from Microsoft Forms).
  - Saves time and reduces transcription errors.

# Types of Data

## Quantitative Data

Numerical data that can be counted or measured.

**Example:** Number of children vaccinated, weight, blood pressure.

- Continuous (Cholesterol levels, blood sugar levels).
- Discrete (Number of children in a family, number of new cases of a disease).

## Categorical Data

Descriptive data that can be categorized.

**Example:** Gender, smoking status (yes/no), type of disease.

- Nominal (Blood type (A, B, AB, O), marital status)
- Ordinal (Stages of cancer (Stage I, II, III, IV)).

### البيانات الكمية: (Quantitative Data)

بيانات رقمية يمكن عدّها أو قياسها (عدد الأطفال الذين تم تطعيمهم، الوزن، ضغط الدم).

#### أنواع البيانات الكمية:

بيانات مستمرة: (Continuous) مثل مستويات الكوليسترول، ومستويات السكر في الدم.

بيانات متقطعة: (Discrete) مثل عدد الأطفال في العائلة، وعدد الحالات الجديدة لمرض معين.

### البيانات النوعية: (Categorical Data)

بيانات وصفية يمكن تصنيفها إلى فئات. (أمثلة: الجنس، حالة التدخين (نعم/لا)، نوع المرض).

#### أنواع البيانات النوعية:

بيانات اسمية: (Nominal) مثل فصيلة الدم (A, B, AB, O)، والحالة الاجتماعية.

بيانات ترتيبية: (Ordinal) مثل مراحل السرطان (المرحلة الأولى، الثانية، الثالثة، الرابعة).

# Data Cleaning

## تنظيف البيانات

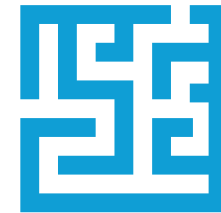
- يتضمن تنظيف البيانات جودة وموثوقية مجموعة البيانات.
- يعتبر خطوة أساسية لتحليل البيانات وتفسيرها بدقة.
- تتضمن الخطوات: معالجة التكرارات، والقيم المفقودة، وإعادة ترميز المتغيرات.



- Data cleaning ensures **quality and reliability** of the dataset.



- Essential for **accurate analysis** and **interpretation**.



- Steps include **handling duplicates, missing values, and recoding variables**.

# Cleaning Data

تنظيف البيانات

1. تحديد وإزالة الإدخالات المكررة.

2. معالجة السجلات ذات البيانات المفقودة بشكل كبير من خلال الحذف أو التعويض.

3. إعادة ترميز المتغيرات لضمان التناسق (مثل تحويل "ذكر" و "M" إلى كود موحد مثل 1 ، وشرح الترميز في دليل البيانات).

4. دمج المتغيرات ذات الصلة (مثل دمج أعمدة "الاسم الأول" و "اسم العائلة").

5. تحويل المتغيرات المستمرة إلى فئات (مثل تحويل "العمر" إلى مجموعات عمرية: 0-18 ، 19-35).

- Identify and remove **duplicate** entries.
- Remove records with excessive **missing** data, or exchange.
- **Recode** variables, for example: transform 'Male' and 'M' into a uniform value (e.g., 1) [1=male, and 2=female]. Then, explain the coding in the data dictionary.
- **Combining** Variables: Merge columns for related data (e.g., 'First Name' and 'Last Name').
- **Converting** Variables: Change continuous variables into categories.
  - Example: Age ? Age Groups (e.g., 0–18, 19–35).



# Ready for Analysis

- After cleaning:
  - Data is accurate and consistent.
  - Free of duplicates and missing values.
  - Variables are appropriately recoded or categorized.
- Next Steps: Import the cleaned dataset into statistical software for analysis.

## الاستعداد للتحليل

- بعد تنظيف البيانات:
- يجب أن تكون البيانات دقيقة ومتناسقة.
- خالية من التكرار والقيم المفقودة.
- المتغيرات مُرمّزة أو مُصنّفة بشكل مناسب.
- الخطوة التالية: استيراد البيانات المُنظفة إلى برامج التحليل الإحصائي لإجراء التحليل.

