الفصل الثالث

ملخص البيانات والإحصاء الوصفي
Data Summary and Descriptive Statistics

نفترض الآن أننا جمعنا بياناتنا من خلال استخدام تصميم تجريبي جيد. لدينا الآن مجتمع إحصائي من الأرقام، أو الملاحظات، أو توصيفات لوصف بياناتنا، ونود تلخيص البيانات لاتخاذ قرارات، أو اختبار فرضية، أو استخلاص نتيجة.

(٣,١) لماذا نجمع البيانات؟
WHY DO WE COLLECT DATA?

العالم مليء بالشك، بمعنى أن هناك عوامل عشوائية أو لا يمكن التنبؤ بها تؤثر على كل قياس تجريبي نقوم به. تنشأ الجوانب غير المتوقعة للنتائج التجريبية أيضاً من المتغيّر في الأنظمة البيولوجية (نتيجة لعوامل وراثية وبيئية) وعمليات التصنيع، والخطأ البشري في إجراء القياسات، والعمليات الأساسية الأخرى التي تؤثر على القياسات التي يتم إجراؤها.

على الرغم من عدم اليقين بشأن النتيجة الدقيقة لتجربة أو حدوث لحادثة في المستقبل، فإننا نجمع البيانات في محاولة لفهم أفضل للعمليات أو المجتمعات الإحصائية التي تؤثر على النتيجة التجريبية حتى نتمكن من القيام ببعض التوقعات. توفّر البيانات معلومات

لتقليل عدم اليقين، والسماح باتخاذ القرارات. وأيضا تجميع وتحليل البيانات يساعدنا في حل المشاكل بشكل صحيح. ولا يمكن التأكيد بما فيه الكفاية على أنه يجب تجميع البيانات وتحليلها بشكل صحيح إذا كان لتحليل البيانات والاستنتاجات اللاحقة أية قيمة.

(٣,٢) لماذا نحن بحاجة إلى الإحصاء؟
WHY DO WE NEED STATISTICS?

لدينا ثلاثة أسباب رئيسية لاستخدام وتحليل البيانات الإحصائية :

١- العالم الحقيقي مليء بالأحداث العشوائية التي لا يمكن وصفها بالصيغ الرياضية الدقيقة.

٢- المتغير هو صفة طبيعية وعادية للعالم الطبيعي.

٣- نود اتخاذ قرارات تتسم ببعض الثقة. وهذا يعني أننا في حاجة إلى إيجاد ميول ضمن المتغير.

(٣,٣) ما هي المسائل التي نأمل معالجتها بالتحليل الإحصائي؟
WHAT QUESTIONS DO WE HOPE TO ADDRESS WITH OUR STATISTICAL ANALYSIS?

هناك العديد من المسائل الأساسية نأمل في معالجتها عند استخدام موجز عددي وبياني للبيانات :

١- هل نستطيع المفاضلة بين المجتمعات الإحصائية أو الجماعات؟

٢- هل هناك ارتباطات بين المتغيرات أو المجتمعات الإحصائية؟

٣- هل العمليات تحت السيطرة؟

إن إيجاد الاختلافات الفسيولوجية بين المجتمعات الإحصائية هو على الأرجح الهدف الأكثر شيوعاً للبحوث الطبية الحيوية. على سبيل المثال، قد يريد الباحثون

معرفة إذا كان هناك اختلاف في متوسط العمر المتوقع بين الناس الذين يعانون من زيادة الوزن والناس الذين يعانون من نقص الوزن. أو قد تريد شركة دوائية تحديد ما إذا كان أحد أنواع المضادات الحيوية أكثر فعالية في مكافحة البكتيريا من نوع آخر. أو يتساءل طبيب عما إذا كان يتم تخفيض ضغط الدم الانبساطي في مجتمع إحصائي من أشخاص يعانون ارتفاع ضغط الدم بعد تناول دواء خافض للضغط. في معظم الأحيان، يقارن باحثو الطبية الحيوية مجتمعات إحصائية من الناس أو الحيوانات خضعت لاثنين أو أكثر من العلاجات المختلفة أو الاختبارات التشخيصية، وهم يريدون معرفة ما إذا كان هناك فرق بين استجابات المجتمعات الإحصائية التي تلقت علاجات أو اختبارات مختلفة. نستخلص في بعض الأحيان عينات متعددة من نفس المجتمع الإحصائي من الأشخاص أو الوحدات التجريبية. وهناك مثال شائع هو عندما يتم أخذ البيانات الفسيولوجية قبل وبعد علاج محدد، مثل امتصاص الأدوية أو العلاج الإلكتروني، من مجتمع إحصائي واحد من المرضى. ندعو هذا النوع من جمع البيانات في التصميم التجريبي "بالحجب". وتم مناقشة مفهوم الحجب هذا بشكل كامل في الفصل الثاني.

إن إحدى المسائل الأخرى التي كثيراً ما تكون هدف البحوث الطبية الحيوية هي ما إذا كان هناك ارتباط بين متغيرين فسيولوجيين. على سبيل المثال، هل هناك ارتباط بين بناء الجسم والوفيات؟ أو هل هناك علاقة بين امتصاص الدهون وحدوث الأورام السرطانية. أو هل هناك ارتباط بين حجم عضلة البطين في القلب وتردد نَظْم القلب غير الطبيعي؟ يتضمن هذا النوع جمع مجموعتين من البيانات وإجراء تحليل الارتباط لتحديد مقدار جودة إمكانية التنبؤ بمجتمع إحصائي من البيانات من الأخرى. وعندما نتكلم عن تحليل الارتباط، فإننا نشير إلى العلاقة الخطية بين متغيرين، والقدرة على التنبؤ بمجتمع إحصائي واحدة من البيانات عن طريق نمذجة البيانات كدالة خطي للمجتمع الإحصائي

الثاني من البيانات. ونظراً لأن تحليل الارتباط يقيس فقط العلاقة الخطية بين عمليتين أو مجموعتين من البيانات، فقد تكون العلاقات غير الخطية بين العمليتين غير واضحة. ويمكن الاطلاع على وصف أكثر تفصيلاً لتحليل الارتباط في الفصل السابع.

وأخيراً، قد يكون المهندس الطبي الحيوي، وخاصة المهندس المشارك في التصنيع، مهتماً في معرفة ما إذا كانت عملية التصنيع تحت السيطرة. قد تنشأ مثل هذه المسائل إذا كانت هناك ضوابط صارمة على مواصفات التصنيع لجهاز طبي. على سبيل المثال، إذا كان المهندس يحاول ضمان جودة إنتاج القثاطر داخل الأوعية الدموية التي يجب أن تكون بأقطار تتراوح بين 1 و 2 سنتيمتر (سم)، فقد يقوم المهندس بجمع عينات من القثاطر عشوائياً من خط التجميع على فترات عشوائية خلال النهار، وقياس أقطارها، وتحديد عدد القثاطر التي تحقق المواصفات، وتحديد ما إذا كان هناك تغيير مفاجئ في عدد القثاطر التي لا تحقق المواصفات. إذا كان مثل هذا التغيير موجودًا، فقد يبحث المهندسون عن عناصر عملية التصنيع التي تتغير مع مرور الزمن، أو المتغيرات في العوامل البيئية، أو أخطاء المُستخدِم. ويمكن للمهندس استخدام مخططات التحكم لتقييم ما إذا كانت العمليات تحت السيطرة. لم يتم تغطية طرق التحليل الإحصائي هذه في هذا الكتاب، ولكن يمكن العثور عليها في عدد من المراجع، بما في ذلك المرجع [3].

(٣,٤) كيف نلخص البيانات بيانيًا؟

HOW DO WE GRAPHICALLY SUMMARIZE DATA?

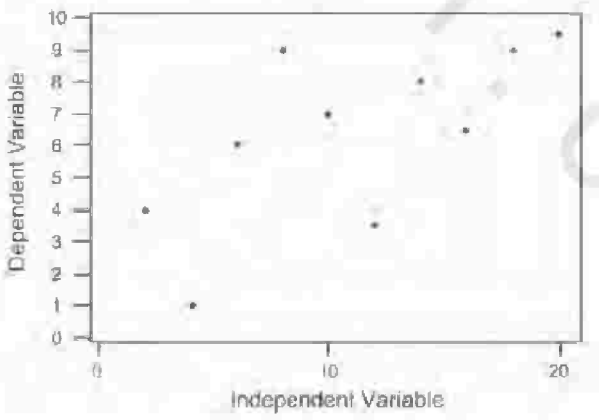
يمكننا تلخيص البيانات بشكل بياني أو عددي. و الصورة العددية هي ما نطلق عليه الإحصائيات. قبل تطبيق التحليل الإحصائي بشكل متهور، فإنه من الجيد دائماً إلقاء نظرة على البيانات الخام، عادة في شكل بياني، ومن ثم استخدام أساليب بيانية لتلخيص البيانات في شكل يسهل فيه تفسيرها.

تشمل أنواع العرض البياني الأكثر استخداماً من قبل المهندسين الطبيين الحيويين ما يلي: المخططات (الرسوم) المبعثرة، والسلاسل الزمنية، المخططات الصندوق والمؤشر، والمدرج التكراري.

إن تفاصيل إنشاء هذه الملخصات البيانية موصوفة في المراجع [3 – 6]، ولكن سوف نصفها هنا بإيجاز.

(٣,٤,١) المخططات المبعثرة Scatterplots

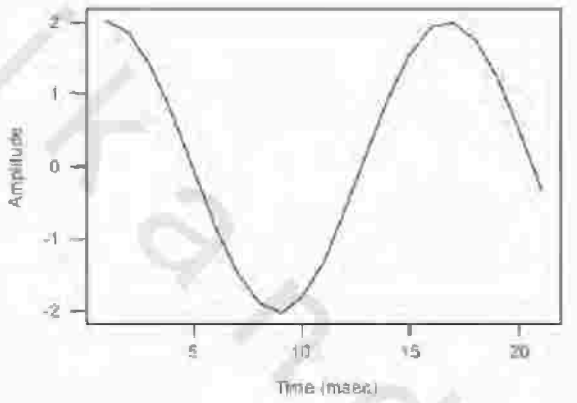
ترسم المخططات المبعثرة ببساطة الحدث لمتغير واحد بالنسبة إلى آخر. في معظم الحالات، يمكن اعتبار أحد المتغيرات المتغير المستقل (مثل الزمن أو رقم الشخص)، ويتم اعتبار المتغير الثاني المتغير التابع. يوضح الشكل (٣,١) مثالاً عن المخططات المبعثرة لمجموعتين من البيانات. بشكل عام، نحن مهتمون فيما إذا كانت هناك علاقة يمكن التنبؤ بها ترسم خريطة المتغير المستقل (مثل معدل التنفس) بالنسبة إلى المتغير الدال لدينا (مثل معدل ضربات القلب). إذا كان هناك علاقة خطية بين المتغيرين، فإنه ينبغي لنقاط البيانات أن تقع قريبة من خط مستقيم.



الشكل (٣,١). مثال على المخطط المبعثر.

(٣,٤,٢) السلاسل الزمنية Time Series

يتم استخدام السلاسل الزمنية لرسم التغييرات في متغير كدالة في الزمن. وعادة ما يكون المتغير مقداراً فسيولوجياً يتغير مع الزمن، مثل النشاط الكهربائي في الدماغ أو تركيز الهرمون في مجرى الدم، . يوضح الشكل (٣,٢) مثالاً على مخطط سلاسل زمنية. وننظر في هذا الشكل إلى دالة الجيب البسيطة عندما يتغير مع الزمن.

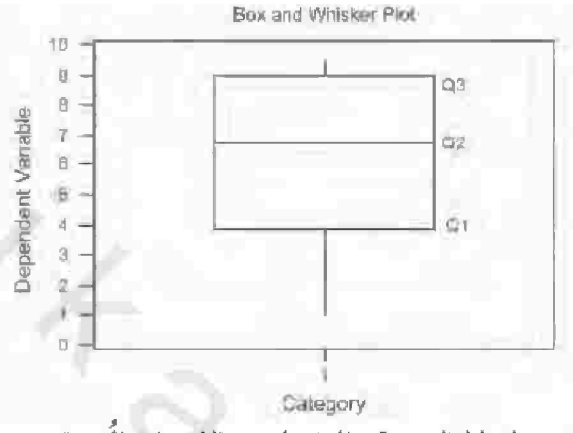


الشكل (٢ , ٣). مثال على مخطط السلاسل الزمنية. يتم رسم سعة العينات كدالة في الزمن.

(٣,٤,٣) المخططات الصندوق والمؤشر Box-and-Whisker Plots Series

توضح هذه المخططات الشرائح الربعية (quartiles) الأولى والثانية والثالثة، بالإضافة إلى القيمة العظمى والقيمة الصغرى لقيم البيانات التي تم جمعها. إن الربع الثاني (Q2) معروف بأنه الوسيط للبيانات. هذه الكمية، على النحو المحدد في وقت لاحق في هذا الكتاب، هي نقطة البيانات الوسطى أو قيمة العينة الوسطى عندما يتم ترتيب العينات في ترتيب تنازلي. يمكن التفكير في شريحة الربيع الأول (Q1) بأنها قيمة متوسط العينات التي تقع تحت شريحة الربيع الثاني. وبالمثل، يمكن التفكير في شريحة الربيع الثالث (Q3) بأنها قيمة متوسط العينات التي تقع فوق شريحة الربيع الثاني.

تُعتبر المخططات الصندوق والمؤشر مفيدة من حيث إنها تبرز ما إذا كان هناك انحراف في البيانات أو أية قيم متطرفة غير عادية في العينات (الشكل ٣.٣).



الشكل (٣,٣). توضيح لمخطط الصندوق والمؤشر لمجتمع الإحصائي المُدرجة. ويتم تبيان شرائح الربيع الأول (Q1) والثاني (Q2) والثالث (Q3). بالإضافة إلى ذلك، تمتد الشعيرات إلى القيمة العظمى والقيمة الصغرى لمجتمع العينة.

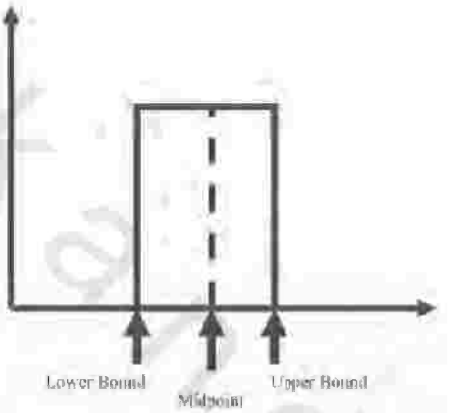
(٣,٤,٤) المدرج التكراري Histogram

يتم تعريف المدرج التكراري بأنه توزيع الترددات. عند الأخذ في الاعتبار N عينات أو قياسات، x_i ، وفى المدى من $X_{\min}$ إلى $X_{\max}$ ، فإنه يتم تجميع العينات في فترات زمنية غير متداخلة (صناديق)، وعادة متساوية العرض (الشكل ٣.٤). يكون عدد الصناديق عادة من 7 – 14، وهذا متوقف على طبيعة البيانات. وبالإضافة إلى ذلك، نتوقع عادة أن يكون لدينا ثلاث عينات على الأقل لكل صندوق [7]. ويمكن أيضاً استخدام قاعدة Sturgess (سترجس) [6] لتقدير عدد الصناديق والتي تُعطى بالعلاقة التالية :

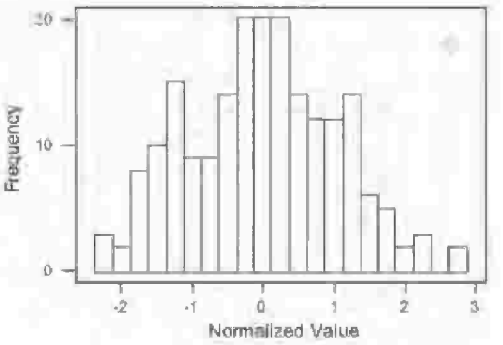
$$k = 1 + 3.3 \log (n)$$

حيث k هو عدد الصناديق و n هو عدد العينات.

كل صندوق من المدرج التكراري لديه حد أدنى، وحد أعلى، ونقطة منتصف. يتم إنشاء المدرج التكراري من خلال رسم عدد العينات في كل صندوق. يوضح الشكل (٣,٥) المدرج التكراري لـ 1000 عينة مأخوذة من توزيع عادي له متوسط (μ)=0 وانحراف معياري (σ)=1.0. لدينا على المحور الأفقي قيمة العينة وعلى المحور العمودي عدد حدوث العينات التي تقع ضمن الصندوق.



الشكل (٣, ٤). أحد صناديق مخطط المدرج التكراري يتم تعريف الصندوق من خلال حد أدنى، ونقطة منتصف، وحد أعلى.



الشكل (٣, ٥). مثال لمخطط مدرج تكراري يتم رسم قيمة المقدار أو العينة على المحور الأفقي في حين يتم رسم تردد حدوث ذلك المقدار أو العينة على طول المحور العمودي.

يوجد مقداران مفيدان في وصف المدرج التكراري هما التردد المطلق والتردد النسبي في واحد أو أكثر من الصناديق. يتم تعريف هذه الكميات كما يلي :

أ) f_i= التردد المطلق في الصندوق رقم i ؛

ب) f/n= التردد النسبي في الصندوق رقم i، حيث n هي العدد الإجمالي للعينات والذي يجري تلخيصه في المدرج التكراري.

هناك عدد من الخوارزميات التي تستخدمها الأجهزة الطبية الحيوية لتشخيص أو كشف التشوهات في وظيفة بيولوجية ؛ وتستفيد هذه الخوارزميات من المدرج التكراري البيانات التي تم جمعها والترددات النسبية المرتبطة بالصناديق المُختارة [8]. وفي كثير من الأحيان يمكن التمييز بين الوظائف الفسيولوجية الطبيعية وغير الطبيعية (أصوات التنفس، وتغيّر معدل ضربات القلب، والمحتوى الترددي للإشارات الفسيولوجية الكهربائية) من خلال مقارنة الترددات النسبية في الصناديق المُستهدفة من المدرج التكراري للبيانات التي تمثل هذه العمليات البيولوجية.

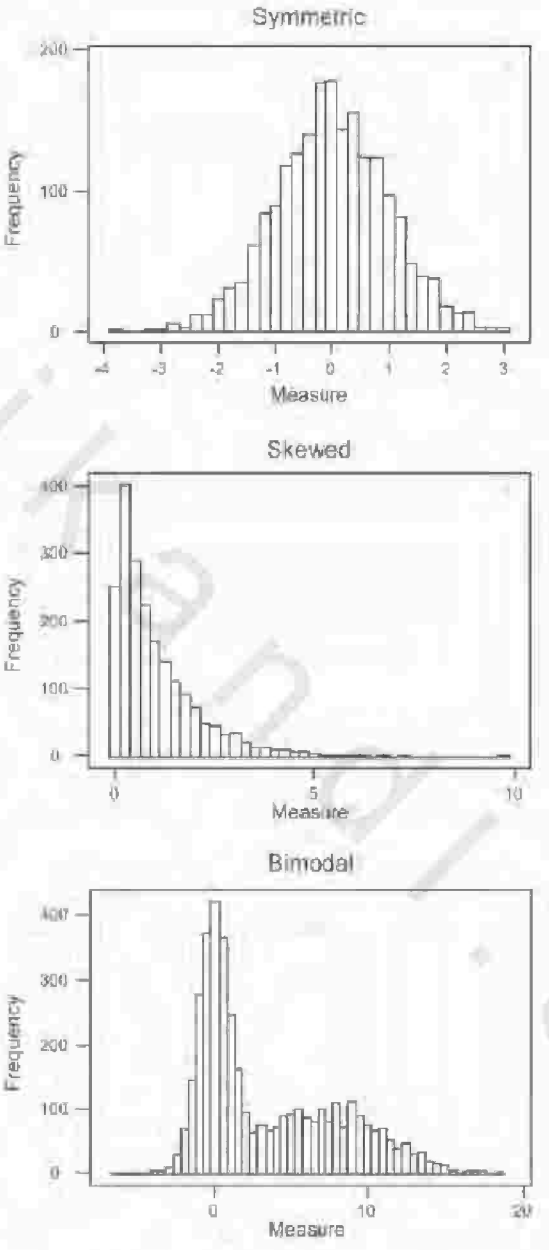
يمكن أن يأخذ المدرج التكراري عدة أشكال. ويُشار إلى الأشكال، المُوضحة في الشكل (٣.٦)، بالمتناظرة، أو غير المتماثلة (المنحرفة)، أو ثنائية الشكل (الدارج).

ويمكن أن يُعزى المدرج التكراري غير المتماثل إلى ما يلي [9] :

١- الآليات ذات الاهتمام التي تولّد البيانات (على سبيل المثال، الآليات الفسيولوجية التي تحدد الفترات الزمنية من ضربة إلى ضربة في القلب).

٢- تشويش صنعي بسبب عملية القياس أو تغيير في الآلية الأساسية مع مرور الزمن (على سبيل المثال، قد يكون هناك تبدلات متغيرة مع الزمن في عملية التصنيع التي تؤدي إلى تغيير في إحصائيات عملية التصنيع مع مرور الزمن).

٣- مزج للمجتمعات الإحصائية التي يتم أخذ العينات منها (وهذا نموذجياً مصدر المدرج التكراري ثنائي الشكل).



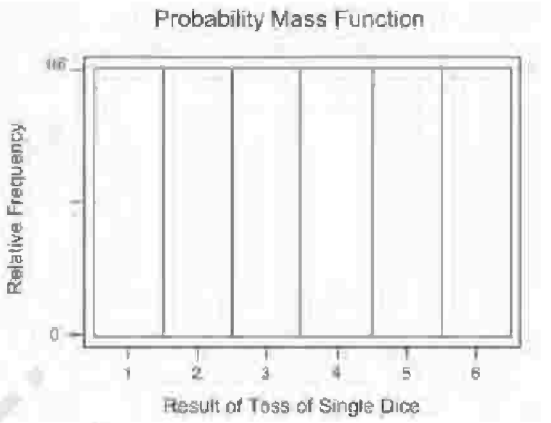
الشكل (٦ , ٣). أمثلة على المدرج التكراري المتناظر (أعلى)، وغير المتماثل (الوسط)، وثنائي الشكل (أسفل).
في كل حالة، تم أخذ 2000 عينة من المجتمعات الإحصائية.

يُعد المدرج التكراري مهمًّا لأنه بمثابة تقدير تقريبي لدالة كثافة الاحتمال الحقيقية أو توزيع الاحتمال للعملية العشوائية الأساسية التي يجري جمع العينات منها.

إن دالة كثافة الاحتمال أو توزيع الاحتمال هي دالة تحدد مقدار احتمال حدوث حادثة عشوائية، x_i. عندما يكون الحدث العشوائي الأساسي متقطعاً في الطبيعة، فإننا نشير إلى دالة كثافة الاحتمال بأنها دالة كتلة الاحتمال [10]. وفي كلتا الحالتين، تصف الدالة الطبيعة الاحتمالية للمتغير أو الحدث العشوائي الأساسي، ويسمح لنا أن نتوقع احتمال ملاحظة نتيجة محددة، x (مُمثَّلة بمتغير عشوائي)، لتجربة ما. إن دالة التوزيع التراكمي هي ببساطة مجموع الاحتمالات لمجتمع إحصائي من النتائج، حيث تكون النتيجة أقل من أو تساوى قيمة محددة لـ x.

دعونا نأخذ في الاعتبار المتغير العشوائي الذي من أجله تكون دالة كثافة الاحتمال مُعرَّفة بشكل جيد (فيما يتعلق بمعظم ظواهر العالم الحقيقي، فإن مثل نماذج الاحتمال هذه تكون غير معروفة). إن المتغير العشوائي هو النتيجة لرمية واحدة للنرد (الزهر). ليكن لدينا نرد واحد واضح ذو ستة أوجه، فإن احتمال تدحرج الرقم ستة من رمية النرد هو 1 من 6. وفي الواقع، فإن احتمال رمي الرقم واحد هو أيضاً 1 من 6. إذا أخذنا بعين الاعتبار جميع النتائج المحتملة لرميات النرد ورسم احتمال ملاحظة أي واحدة من تلك النتائج الستة في رمية واحدة، سيكون لدينا مخطط مثل ذلك المُبيّن في الشكل (٣,٧).

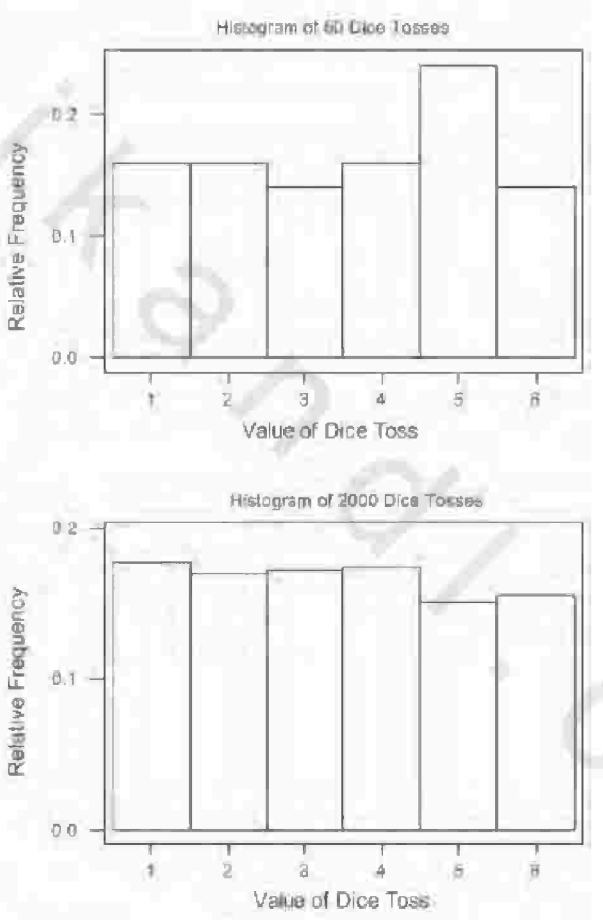
يبين هذا المخطط دالة كثافة الاحتمال أو دالة كتلة الاحتمال لرمية من النرد. ويُعرف هذا النوع من نماذج الاحتمال بالتوزيع المنتظم لأن كل نتيجة لديها نفس احتمال الحدوث بالضبط (1/6 في هذه الحالة).



الشكل (٣,٧). دالة كثافة الاحتمال لمتغير عشوائي متقطع (دالة كتلة الاحتمال). في هذه الحالة، فإن المتغير العشوائي هو قيمة الرمية لنرد (زهر) واحد. لاحظ أن كل نتيجة من النتائج الست المحتملة لديها احتمال حدوث مقداره 1 من 6. إن دالة كثافة الاحتمال هذا معروفة أيضاً بأنها توزيع الاحتمال المنتظم.

نحن نعلم توزيع الاحتمال الحقيقي بالنسبة لرمية النرد. ومع ذلك، بالنسبة لمعظم العمليات العشوائية في العالم الحقيقي، وخاصة العمليات البيولوجية، فإننا لا نعرف كيف تبدو دالة كثافة الاحتمال الحقيقي أو دالة كتلة الاحتمال. ونتيجة لذلك، علينا استخدام المدرج التكراري، الذي يتم إنشاؤه من عينة صغيرة، في محاولة لتقدير التوزيع الأفضل للاحتمال أو نموذج الاحتمال لوصف الظاهرة في العالم الحقيقي. إذا عدنا إلى مثال رمية النرد، فإنه يمكننا في الواقع رمي النرد عدة مرات ورؤية مقدار تطابق المدرج التكراري، الذي تم الحصول عليه من البيانات التجريبية، مع دالة كتلة الاحتمال الحقيقي للنرد المثالي سداسي الوجوه. يوضح الشكل (٣,٨) المدرجات التكرارية لنتائج رميات عددها 50 و 1000 رمية لنرد واحد. لاحظ أنه حتى مع 50 رمية أو عينة، فإنه من الصعب تحديد كيف يمكن أن يبدو توزيع الاحتمال الحقيقي. ومع ذلك، عندما نقترب من 1000 عينة، فإن المدرج التكراري يقترب من دالة كتلة الاحتمال الحقيقي (توزيع منتظم) لرمية النرد. ولكن، لا يزال هناك بعض التباين من صندوق إلى صندوق لذلك لا يبدو منتظماً مثل توزيع الاحتمال

المثالي المُوضح في الشكل (٣.٧). إن الرسالة التي يمكن قراءتها من هذا التوضيح هي أن معظم الأبحاث الطبية الحيوية تنشر النتائج لعدد قليل من العينات. ويتضح من مثال النرد أنه من الصعب جدًّا تحديد الإحصائيات للعملية العشوائية الأساسية من عينة صغيرة، ولكن معظم الأبحاث الطبية الحيوية تعتمد على بيانات من عينات صغيرة.



الشكل (٣ ,٨). المدرجات التكرارية تمثل نتائج التجارب التي يتم فيها رمي نرد واحد 50 مرة (أعلى) و 2000 مرة (أسفل)، على التوالي لاحظ أنه كلما ازداد حجم العينة، اقترب المدرج التكراري من توزيع الاحتمال الحقيقي المُوضح في الشكل (٣.٧).

(٣,٥) النهج العام للتحليل الإحصائي
GENERAL APPROACH TO STATISTICAL ANALYSIS

لقد جمعنا بياناتنا الآن وألقينا النظر على بعض الملخصات البيانية للبيانات. والآن سوف نستخدم ملخصاً عدديًا، معروفاً أيضاً باسم الإحصائيات، في محاولة لوصف طبيعة المجتمع الإحصائي أو العملية التي أخذنا منها العينات. نفترض، من هذه الإحصائيات الوصفية، نموذج احتمال أو توزيع احتمال للمجتمع الإحصائي أو العملية ومن ثم نختار الاختبارات الإحصائية المناسبة لاختبار الفرضيات أو اتخاذ القرارات. ومن المهم ملاحظة أن الاستنتاجات التي يمكن للمرء استخلاصها من تجربة إحصائية تعتمد على مدى ملاءمة نموذج الاحتمال المُفترض للمجتمع الإحصائي أو العملية.

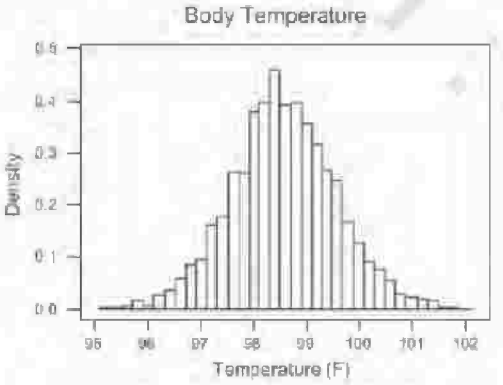
وكما جاء في المقدمة، يحاول مهندسو الطبية الحيوية اتخاذ قرارات حول المجتمعات الإحصائية أو العمليات التي لديهم وصول محدود إليها. وهكذا، فهم يصممون التجارب ويجمعون العينات التي يعتقدون أنها إلى حد ما تمثل المجتمع الإحصائي أو العملية. وبغض النظر عن نوع التحليل الإحصائي الذي سوف ينتج عن البحث أو الدراسة، فإنه ينبغي لجميع أنواع التحليل الإحصائي أن تتبع نفس النهج العام:

١- قياس عدد محدود من عينات تمثيلية من المجتمع الإحصائي الأكبر.

٢- تقدير الإحصاءات الحقيقية للمجتمع الإحصائي الأكبر من إحصائيات العينة.

يجب هنا معالجة بعض المفاهيم الهامة. المفهوم الأول واضح إلى حد ما. غالباً ما يكون مستحيلاً أو غير عملي أخذ القياسات أو الملاحظات من المجتمع الإحصائي الكامل. وهكذا، فإن المهندس الطبي الحيوي عادة ما يختار عينة أصغر وأكثر عملية تمثل المجتمع الإحصائي ومدى المتغير في المجتمع الإحصائي الأكبر. على سبيل المثال، قد لا نستطيع قياس درجة حرارة الجسم في حالة الراحة لكل إنسان على وجه الأرض للحصول على تقدير لدرجة حرارة الجسم الطبيعية والمجال الطبيعي لها. يتركز اهتمامنا هنا على معرفة ما هي درجة حرارة الجسم الطبيعية، في المتوسط،

للإنسان السليم والمجال الطبيعي لدرجة الحرارة في حالة الراحة بالإضافة إلى احتمال أو إمكانية قياس درجة حرارة الجسم الخاصة في الحالات الصحية، وفي ظروف الراحة. في محاولة لتحديد الخصائص أو نموذج الاحتمال الأساسي لدرجة حرارة الجسم للأفراد الأصحاء، وفي حالة الراحة، سيختار الباحث، على نحو عشوائي، عينة من الأفراد الأصحاء وفي حالة الراحة ويقيس درجة حرارتهم الخاصة في حالة الراحة بواسطة مقياس حرارة. يجب أن يراعي الباحثون تكوين وحجم مجتمع العينة لتمثيل المتغير بشكل كاف في المجتمع الإحصائي. ويجب على الباحث تحديد ما يميّز الفرد الطبيعي والسليم، مثل العمر والحجم والعرق والجنس والصفات الأخرى. إذا كان الباحث يجمع بيانات درجة حرارة الجسم من عينة مؤلفة من 3000 شخص مثلاً، فقد يرسم المدرج تكرارياً لدرجات الحرارة التي تم قياسها من الـ 3000 شخص ويخلص في النهاية إلى المدرج التكراري التالي المُبين في الشكل (٣,٩). وقد يحسب الباحث أيضاً بعض الإحصائيات الوصفية الأساسية لـ 3000 عينة، مثل متوسط العينة (الوسيط الحسابي) (average (mean)، الوسيط (median) (القيمة الواقعة في المنتصف)، والانحراف المعياري (standard deviation).



الشكل (٣,٩). المدرج التكراري لـ 2000 قيمة لدرجة حرارة الجسم الداخلية تم جمعها من المجتمع الإحصائي مُوزَّعة بشكل طبيعي.

يحاول الباحث، بعد تقدير إحصائيات العينة من مجتمع إحصائي، استخلاص استنتاجات حول المجتمع الإحصائي الأكبر (الحقيقية). إن السؤال الأكثر أهمية الذي يجب طرحه عند مراجعة الإحصائيات والاستنتاجات المُستخلصة من مجتمع إحصائي هو مقدار جودة تمثيل مجتمع العينة للمجتمع الإحصائي الأكبر.

نستخدم، بعد جمع البيانات، بعض الإحصائيات الوصفية الأساسية لتلخيص البيانات. وتشمل هذه الإحصائيات الوصفية الأساسية المقادير العامة التالية: النزعة المركزية، والمتغير، والارتباط.

(٣,٦) الإحصاء الوصفي
DESCRIPTIVE STATISTICS

هناك عدد من الإحصائيات الوصفية التي تساعدنا على تصوّر توزيع المجتمع الإحصائي. وبعبارة أخرى، إن هدفنا النهائي هو افتراض نموذج احتمال أساسي للمجتمع الإحصائي ومن ثم اختيار التحليلات الإحصائية المناسبة لنموذج الاحتمال هذا.

عندما نحاول استخلاص استنتاجات حول المجتمع الإحصائي الأكبر أو العملية من العينة الأصغر للبيانات لدينا، فإننا نفترض أن النموذج الأساسي لأي عينة، أو "حدث"، أو مقدار (نتيجة التجربة) هو كما يلي:

$$X = \mu \pm \text{الفروق الفردية} \pm \text{العوامل الظرفية} \pm \text{المتغيرات غير المعروفة}.$$

حيث X هو المقدار أو قيمة العينة لدينا ويتأثر بـ μ، الذي هو متوسط المجتمع الإحصائي الحقيقي؛ الفروق الفردية مثل التركيب الوراثي، والتدريب، والدافع، والحالة الفيزيائية؛ والعوامل الظرفية مثل العوامل البيئية؛ والمتغيرات غير المعروفة مثل العوامل غير المعروفة / العوامل غير المُقاسة التي تتصرف بطريقة لا يمكن التنبؤ بها من لحظة إلى لحظة.

وبعبارة أخرى، عندما نقوم بإجراء قياس أو ملاحظة، فإن القيمة المُقاسة تمثل أو تكون متأثرة ليس فقط بإحصائيات المجتمع الإحصائي، مثل متوسط المجتمع الإحصائي، ولكن بعوامل مثل المتغير البيولوجي من فرد إلى فرد، والعوامل البيئية (الزمن، ودرجة الحرارة، والرطوبة، والإضاءة، والأدوية، وما إلى ذلك)، والعوامل العشوائية التي لا يمكن التنبؤ بها تماماً من لحظة إلى لحظة. تؤدي جميع هذه العوامل إلى إنشاء المدرج التكراري لبيانات العينة، التي قد تعكس أو لا تعكس دالة كثافة الاحتمال الحقيقي للمجتمع الإحصائي. إذا قمنا بعمل جيد في التصميم التجريبي وجمعنا عددًا كافيًا من العينات، فإنه ينبغي للمدرج التكراري والإحصائيات الوصفية لمجتمع العينة أن تعكس بشكل متقارب دالة كثافة الاحتمال الحقيقي والإحصائيات الوصفية للمجتمع الإحصائي الحقيقي. إذا كان هذا هو الحال، يمكننا بعد ذلك القيام باستنتاجات حول المجتمع الإحصائي الأكبر من مجتمع العينة الأصغر. وإذا كان مجتمع العينة لا يعكس تغيّر المجتمع الإحصائي الحقيقي، فقد تكون عندئذ الاستنتاجات التي قمنا باستخلاصها من التحليل الإحصائي لبيانات العينة ذات قيمة صغيرة.

هناك عدد من نماذج الاحتمال المفيدة لوصف العمليات البيولوجية وعمليات التصنيع. وتشمل التوزيع الطبيعي، وتوزيع بواسون، وتوزيع الأسس، وتوزيع جاما [10]. سنركز في هذا الكتاب على المجتمعات الإحصائية التي تتبع توزيعاً طبيعياً لأن هذا هو توزيع الاحتمال، الأكثر تكراراً والذي تتم مصادفته، المُستخدم في وصف المجتمع الإحصائي. وعلاوة على ذلك، تفترض الطرق الأكثر استخداماً للتحليل الإحصائي أن تكون البيانات مُنمذجة جيداً بواسطة التوزيع الطبيعي ("منحنى الجرس"). ومن المهم ملاحظة أن العديد من العمليات البيولوجية غير مُنمذجة بشكل جيد بواسطة التوزيع الطبيعي (مثل تغيّر معدل ضربات القلب)، والإحصائيات المرتبطة بالتوزيع

الطبيعي ليست مناسبة لمثل هذه العمليات. في مثل هذه الحالات، قد تخدم الإحصائيات غير البارامترية (nonparametric)، التي لا تفترض نوعاً محدداً لتوزيع البيانات، الباحث بشكل أفضل في فهم العمليات واتخاذ القرارات. ومع ذلك، فإن استخدام التوزيع الطبيعي والإحصائيات المرتبطة به غالباً ما يكون كافياً باعتبار نظرية النزعة المركزية، التي تنص ببساطة على أن جمع العمليات العشوائية ذات التوزيع العشوائي سينتج عنه متغير عشوائي ذو توزيع طبيعي. يمكن للمرء الافتراض أن معظم الظواهر البيولوجية تنتج من جمع العمليات العشوائية.

(٣,٦,١) مقاييس النزعة المركزية Measures of Central Tendency

هناك العديد من المقاييس التي تعكس النزعة المركزية أو التركيز لمجتمع العينة: متوسط العينة (المتوسط الحسابي)، وسيط (منتصف) العينة، ومنوال العينة.

يمكن تقدير متوسط العينة من مجتمع من العينات، x_i، حيث i هو رقم العينة، باستخدام الصيغة أدناه.

مُعطى n نقطة من البيانات، $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

في الواقع نحن لا نعرف عادة المتوسط الحقيقي، μ، للمجتمع الإحصائي، بدلاً من ذلك نحاول تقدير المتوسط الحقيقي، μ، للمجتمع الإحصائي الأكبر. عندما يصبح حجم العينة كبيراً، ينبغي لمتوسط العينة، $\bar{x}$، أن يقترب من المتوسط الحقيقي، μ، على افتراض أن إحصائيات المجتمع الإحصائي أو العملية لا تتغير بمرور الزمن أو بتغير المكان.

إن إحدى المشاكل التي تنشأ عند استخدام متوسط العينة لتمثيل النزعة المركزية للمجتمع الإحصائي هي أن متوسط العينة عرضة للقيم المتطرفة. يمكن أن يكون هذا مشكلة وغالباً مضللاً عند وصف متوسط المجتمع الإحصائي المنحرف بشكل كبير.

على سبيل المثال، عند وصف الدخل لمجتمع من خريجي الجامعات الجدد التي تضم لاعب كرة سلة (NBA) وقع عقداً بعدة ملايين من الدولارات، فإن متوسط الدخل المُقدَّر سوف يكون أكبر بكثير مما يكسب معظم الخريجين. إن التحريف (سوء التمثيل) نفسه واضح في كثير من الأحيان عند وصف قيمة المتوسط للمنازل في منطقة جغرافية محددة حيث يمكن لعدد قليل من المنازل قيمتها في حدود المليون إخفاء حقيقة أن عدة مئات من المنازل الأخرى قيمتها أقل من 200000 دولار أمريكي.

ثمة مقياس آخر مفيد لتلخيص النزعة المركزية للمجتمع الإحصائي هو وسيط العينة. إن قيمة الوسيط لمجتمع إحصائي من الملاحظات أو العينات، x_i ، هو الملاحظة الوسطى عندما يتم ترتيب العينات، x_i ، في ترتيب تنازلي.

على سبيل المثال، إذا كان لدينا القيم التالية للحجم المدّي للرئة:

2, 1.5, 1.3, 1.8, 2.2, 2.5, 1.4, 1.3,

فإنه يمكننا إيجاد قيمة الوسيط أولاً من خلال ترتيب البيانات ترتيبًا تنازليًّا:

2.5, 2.2, 2.0, 1.8, 1.5, 1.4, 1.3, 1.3,

بعد ذلك نشطب القيم على كل نهاية حتى نصل إلى قيمة المنتصف:

~~2.5~~, ~~2.2~~, ~~2.0~~, 1.8, 1.5, ~~1.4~~, ~~1.3~~, ~~1.3~~,

في هذه الحالة، هناك قيمتان للوسيط، وهكذا، فإن الوسيط هو متوسط كلتا القيمتين، وهو 1.65.

لاحظ أنه إذا كان عدد العينات، n، عددًا فرديًّا، فإن الوسيط سيكون الملاحظة الوسطى. وإذا كان حجم العينة، n، عددًا زوجيًّا، فإن الوسيط عندئذ يساوي متوسط الملاحظتين الموجودتين في المنتصف. إن وسيط العينة أقل عرضة للتطرف مقارنة مع متوسط العينة. إنه يتجاهل الانحراف في مجتمع إحصائي من العينات أو في دالة كثافة

الاحتمال للمجتمع الإحصائي. بشكل عام، لتمثيل النزعة المركزية بوضوح لمجتمع من العينات أو المجتمع الإحصائي، فإننا نستخدم القاعدة التالية للإبهام:

١- إذا كان المدرج التكراري للعينة أو دالة كثافة الاحتمال للمجتمع الإحصائي متماثلا، استخدم المتوسط كمقياس مركزي. بالنسبة لمثل هذه المجتمعات الإحصائية، فإن المتوسط والوسيط متساويين تقريباً، وتقدير المتوسط يستخدم كافة البيانات.

٢- إذا كان المدرج التكراري للعينة أو دالة كثافة الاحتمال للمجتمع الإحصائي غير متماثل، فإن الوسيط هو المقياس الأكثر ملاءمة لمركز التوزيع.

المقياس الآخر للنزعة المركزية هو المنوال، الذي هو ببساطة الملاحظة الأكثر تكراراً في مجتمع العينات. في مثال الحجم المدّي المذكورة سابقاً، فإن العدد 1.3 هو قيمة العينة الأكثر تكراراً. لا يتم استخدام المنوال بنفس تكرار استخدام المتوسط أو الوسيط في تمثيل النزعة المركزية.

(٣,٦,٢) مقاييس المتغير Measures of Variability

تعتبر مقاييس النزعة المركزية وحدها غير كافية لتمثيل إحصائيات المجتمع الإحصائي أو العملية. من المألوف في الواقع وجود اختلاف في المجتمع الإحصائي الذي يجعل الأمور مثيرة للاهتمام ويؤدي إلى عدم اليقين في صنع القرار. إن الاختلاف من شخص إلى شخص، وخصوصاً في وظيفة فسيولوجية، هو ما يجعل إيجاد التشخيص والعلاج غير المضلل في كثير من الأحيان صعباً للغاية. إن ما يصلح لشخص واحد غالباً ما يفشل مع شخص آخر، وليس المتوسط أو الوسيط هو الذي يتم الحصول عليه من تلك الاختلافات من شخص إلى شخص، بل هو الاختلاف، الذي يتم عكسه في اختلاف نماذج الاحتمال التي تنطوي تحتها تلك المجتمعات الإحصائية المختلفة.

عند تلخيص اختلاف مجتمع أو عملية، فإننا نسأل عادة، "ما هو بعد موقع العينات (البيانات) عن المركز (متوسط العينة)؟". للإجابة على هذا السؤال، نستخدم

عادة التقديرات التالية التي تمثل انتشار بيانات النموذج: مدى الربيعى، تباين العينة، والانحراف المعياري للعينة.

إن مدى الربيعى هو الفرق بين الربيع الأول والثالث لبيانات العينة. وبالنسبة للبيانات التي تم أخذ عينات لها، فإن الوسيط معروف أيضاً بأنه الربيع الثاني، Q2. وبالأخذ في الاعتبار Q2، يمكننا إيجاد الربيع الأول، Q1، ببساطة من خلال أخذ قيمة الوسيط لتلك العينات التي تقع تحت الربيع الثاني. يمكننا إيجاد الربيع الثالث، Q3، بأخذ قيمة الوسيط لتلك العينات التي تقع فوق الربيع الثاني. وكتوضيح لذلك، لدينا العينات التالية:

1, 3, 3, 2, 5, 1, 1, 4, 3, 2,

إذا رتبنا هذه العينات ترتيباً تنازلياً،

5, 4, 3, 3, 3, 2, 2, 1, 1, 1,

إن قيمة الوسيط والربيع الثاني لهذه العينات هي 2.5. ويمكن إيجاد الربيع الأول، Q1، من خلال إيجاد وسط العينات التالية،

2.5, 2, 2, 1, 1, 1,

الذي يساوي 1.5. بالإضافة إلى ذلك، قد يتم إيجاد الربيع الثالث، Q3، بأخذ قيمة وسط العينات التالية:

5, 4, 3, 3, 3, 2.5,

الذي يساوي 3. وهكذا، فإن مدى الربيعي، 1.5 = 3 -1.5 = Q3 – Q1.

يتم تعريف تباين العينة، s^2، بأنه "متوسط المسافة للبيانات عن المتوسط"، وصيغة تقدير s^2 من مجتمع من العينات، x_i، هو:

$$s^2 = \frac{1}{n-1}\sum_{i-1}^{n}\left(x_i - \overline{x}\right)^2$$

إن الانحراف المعياري للعينة، s، الذي غالباً ما يُشار إليه في وصف تغيّر البيانات هو:

$s=\sqrt{s^2}$ (نفس وحدات العينات الأصلية).

من المهم الإشارة إلى أن التوزيعات الطبيعية (المدرجات التكرارية المتماثلة)، ومتوسط العينة وانحراف العينة هي البارامترات الوحيدة اللازمة لوصف إحصائيات الظاهرة الأساسية. وهكذا، إذا كان ينبغي للمرء مقارنة مجموعتين أو أكثر موزعتين طبيعياً، فإنه يحتاج فقط إلى اختبار تكافؤ المتوسطات والتباين لتلك المجتمعات الإحصائية.

obeikandl.com