الورشة التدريبية

مبادئ التحليل الإحصائي للبيانات

د. عيسى المصاروه

المعهد العربي للتدريب والبحوث الإحصائية

عمان - الأردن

الغاية من الورشة

- تعريف المشاركين بالأساليب الإحصائية الأساسية المستخدمة في مجال تحليل البيانات
- تمكين المشاركين من انتاج المؤشرات الإحصائية وتفسيرها واستخدامها في مجالات عملهم

لا بد من استعمال الإحصاءات الرقمية والبيانية
عند تناول كافة النشاطات الوطنية
خاصة الاقتصادية

أشكال المعلومات الرقمية

- البيانات الخام – ضرورية ولكن ؟ أمثلة
- الإحصاءات التلخيصية:
  - الأرقام المطلقة
  - الجداول
  - النسب
  - النسب من واحد صحيح والنسب المئوية (%)
  - نسب التغير
  - المعدلات (للواقعات أو المناسبات)
  - مقاييس النزعة المركزية مثل المنوال والوسط والوسيط
  - مقاييس التشتت
  - الأشكال البيانية المختلفة (يمكن إضافة إحصاءات رقمية إليها ايضاً)
  - الصور المعززة بالأرقام

المعلومات – لماذا ؟

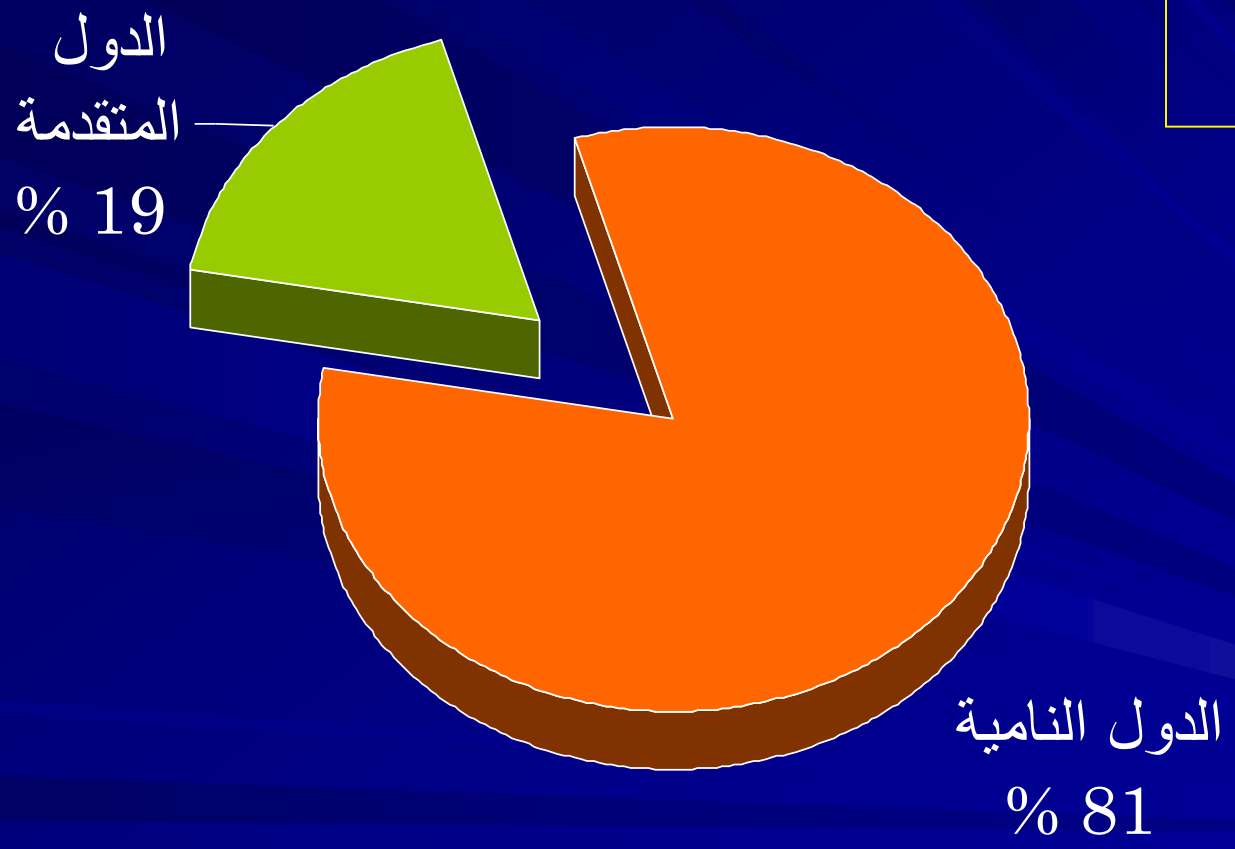
1) للتشخيص - لفهم الوضع الراهن واكتشاف القضايا والاحتياجات ومن ثم الاستجابة لها

2) لاستشراف المستقبل وتوقعه مسبقاً

3) للتقييم - لتقدير التغيرات والنجاحات التي تحققت خلال فترة زمنية معينة سابقة

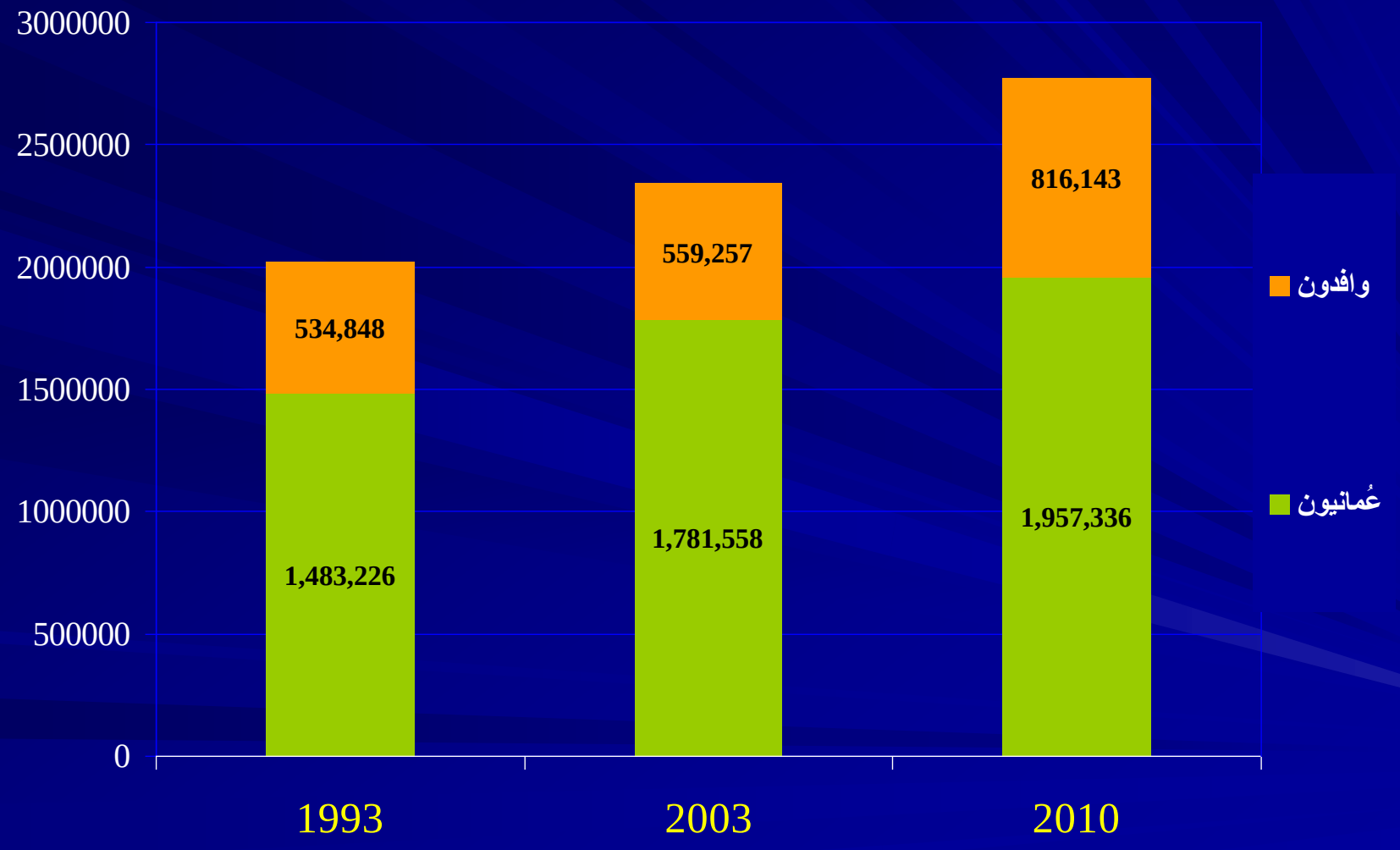
التشخيص - حالة سكان العالم اليوم

7 بليون



Source: 2011 PRB World Population Data Sheet

التقييم - التغير في عدد سكان سلطنة عُمان (مليون)



Source : Oman Censuses

أساسيات في التحليل الإحصائي

- المتغيرات - إذهب إلى

- StatTrek.com
  - AP Statistics - Exploring Data
    - The basics
      - Variables

أساسيات في التحليل الإحصائي

- مستويات قياس المتغيرات - إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - Regression
      - Measurement Scales

المتغيرات – تعريفها ومستويات قياسها

- المتغير هو حالة أو ظاهرة أو سلوكات تتغير وتتباين في الزمان والمكان

  - مثال كمية الأمطار التي سقطت في سلطنة عمان في الفترة 2006-2011
  - كمية الأمطار التي سقطت في اليمن وعُمان عام 2011

- مستويات قياسها :
  - المستوى الأسمي – مثل نوع الاسماك
  - المستوى الترتيبي – دخل عالي جداً ، عالي ، متوسط ، دون المتوسط ، متدني
  - المستوى الفتري – درجات الحرارة الفهرنهاتية ، مقياس الذكاء
  - المستوى النسبي – كميات الأسماك التي تم صيدها ، كميات التمر أو اللُبان التي تم انتاجها

تحليل البيانات الخام أي تلخيصها

1) الأساليب أو الطرق الرقمية
   - الأرقام المطلقة
   - الجداول التكرارية
   - النسب
   - المعدلات
   - مقاييس النزعة المركزية
2) الأساليب أو الطرق البيانية
   - المضلع التكراراي - الخط أو المنحنى المتكسر (الصاعد والنازل)
   - المدرج التكراري – الأعمدة الأفقية والرأسية
   - الدائرة
   - السطح المئوي

تحليل البيانات الخام أي تلخيصها

- الجداول- إذهب إلى

- StatTrek.com
  - AP Statistics - Exploring Data
    - Categorical data
      - One-way table
      - Two-way table

الجداول التكرارية وقواعد تصميمها

1) إعداد الجداول التكرارية: عندما يكون عدد المفردات المراد توحيدها وعرضها ضمن جدول تكراري محدداً، من الممكن أن نعمد إلى الخطوات التالية:

- تحديد الفئات التي يراد التوحيد على أساسها
- وضع جدول تفريغ عددي يتلائم مع الفئات المحددة
- تفريغ المعلومات في الجداول بحيث توضع اشارة مائلة متوازية في خانة التفريغ المناسبة

وكلما تجمع لدينا خمس اشارات في خانة ما، نشطب بالاشارة الخامسة على الاشارات الأربع السابقة

- جمع عدد الإشارات المقابلة لكل فئة فنحصل على جدول تكراري

أما إذا كان عدد المعلومات كبيراً، فيصار إلى إعطاء الخواص المدروسة رموزاً رقمية ومن ثم تُفرغ هذه الرموز الرقمية إما بواسطة جدول التفريغ العددي أو آلياً

تابع الجداول .....

2) الفئات: الفئة هي حدان أو مدى ضمنه مجموعة من المفردات التي تؤلف فيما بينها مجموعة متقاربة من حيث الصفة المدروسة

- عدد الفئات: من المستحسن أن لايقل عدد الفئات عن خمس فئات وأن لا يزيد عن عشرين ويحدد عددها بقسمة المدى العام على طول الفئة
- مدى الفئة أو طولها: يحدد مدى الفئة بقسمة المدى العام على عدد الفئات أو بالاستناد إلى قانون "ستورجز"

طول الفئة = المدى ÷ لغ ن (3.322) + 1

في المثال: المدى = 35 – 16 = 19

لغ ن = لغ 112 = 2.04922

طول الفئة = 19 ÷ (3.322 × 2.04922) + 1

19 ÷ 7.81 = 2.43

طول الفئة = 2.43 وبالتقريب = 2

تابع الجداول .....

3) عناوين أعمدة الجدول
4) عناوين أسطر الجدول
5) مجاميع الأعمدة ومجاميع الأسطر
6) المجموع العام
7) عنوان الجدول – ما هي صفاته؟
8) عدد خلايا الجدول = عدد الأعمدة × عدد الأسطر (باستثناء عمود وسطر مجاميع الأعمدة والأسطر)
9) مصدر المعلومات
10) ما عدد المتغيرات التي يمكن وضعها في الجدول ؟
11) إذا كان لدي متغيران، فأيهما أضعه في الأعمدة ؟
12) ما هي الإحصاءات التي أضعها في الجدول ؟

التكرار المتجمع

لكثير من الأغراض الإحصائية يتم جمع التكرارات من الأعلى إلى الأسفل وهو ما يعرف بالتكرار المتجمع الصاعد، كما يتم جمعها من الأسفل إلى الأعلى وهو مايعرف بالتكرار المتجمع النازل (أو الهابط)

التوزيع التكراري المزدوج

- هناك بيانات مزدوجة تتكون من متغيرين مثلا (مساحة المساكن وعدد الغرف في كل منها)
- حين يكون عدد الارقام أكبر يفضل توزيعها توزيعا تكراريا يعرف بالتوزيع التكراري المزدوج
- حيث تتكون لدينا مجموعتين من الفئات توضع احدهما بالاتجاه الافقي وتوضع الأخرى بالاتجاه الرأسي
- توزع أزواج الأرقام عليها

تطبيقات عملية على الجداول

- عمل جدول تكراري
  - لمتغير واحد (نوع الأسماك)
  - لمتغيرين معاً (نوع الأسماك ووزنها)
- حسب القواعد المتعارف عليها
- استخدم البيانات عن نوع الأسماك وأوزانها

عرض المعلومات بيانياً

- إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - Charts and graphs
      - Patterns in data
      - Dot plots
      - Histogram
      - Stem plots
      - Cumulative plots
      - Scatter plots
      - Compare data sets

عرض التوزيع التكراري بيانياً

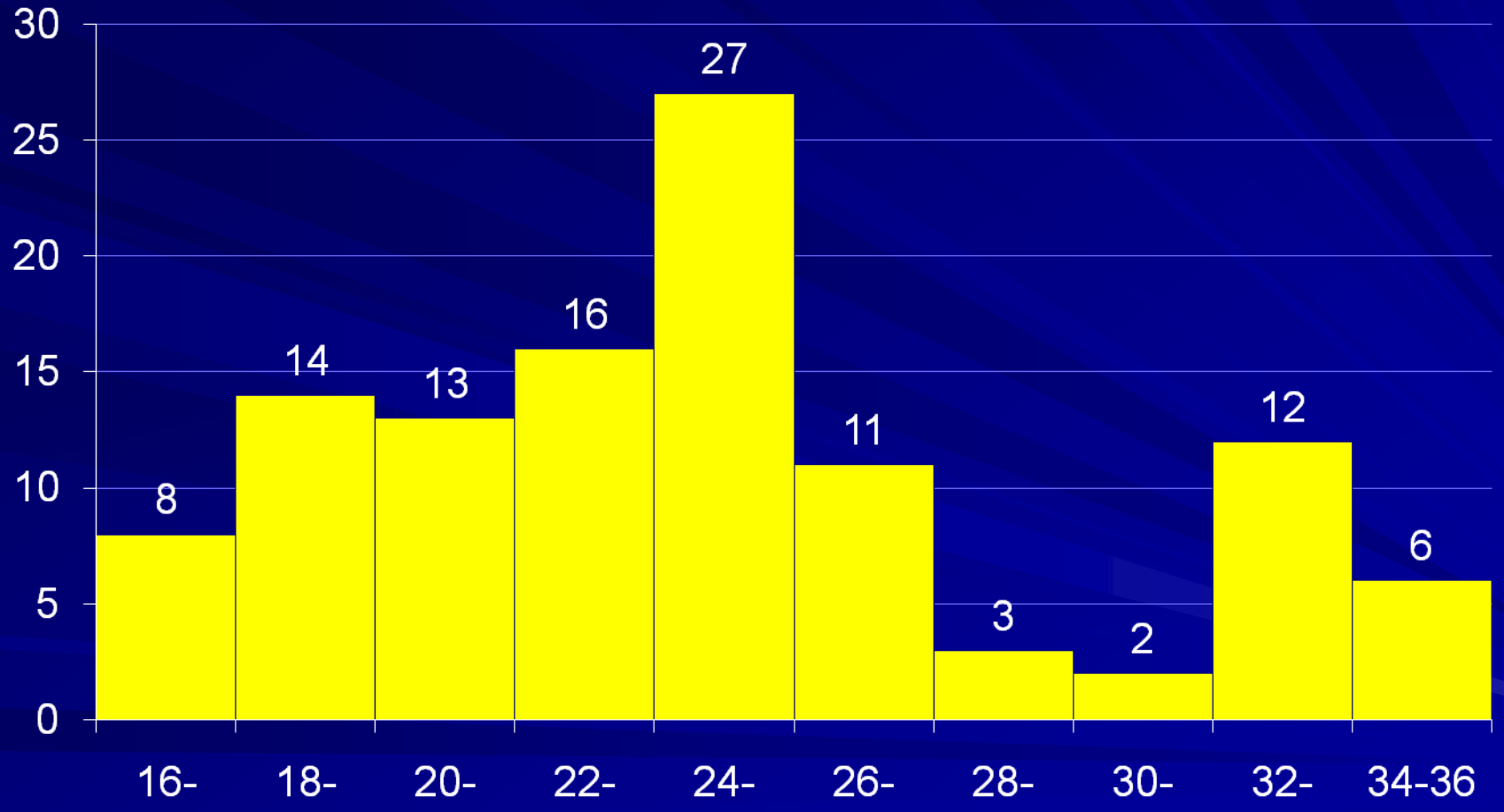
من الممكن عرض المعلومات الواردة في جدول تكراري بيانياً، إما بواسطة الأعمدة (المدرج التكراري) أو بواسطة المنحنى المنكسر (المضلع) أو بواسطة المنحنى الممهد

أ) المدرج التكراري أو الهيستوغرام: يمكننا أن نعرض معلومات التوزيع التكراري بيانياً وذلك بإعطاء كل فئة عموداً مستقلاً طول هذا العمود بمقدار تكررات الفئة نفسها والشكل التالي يوضح الفكرة

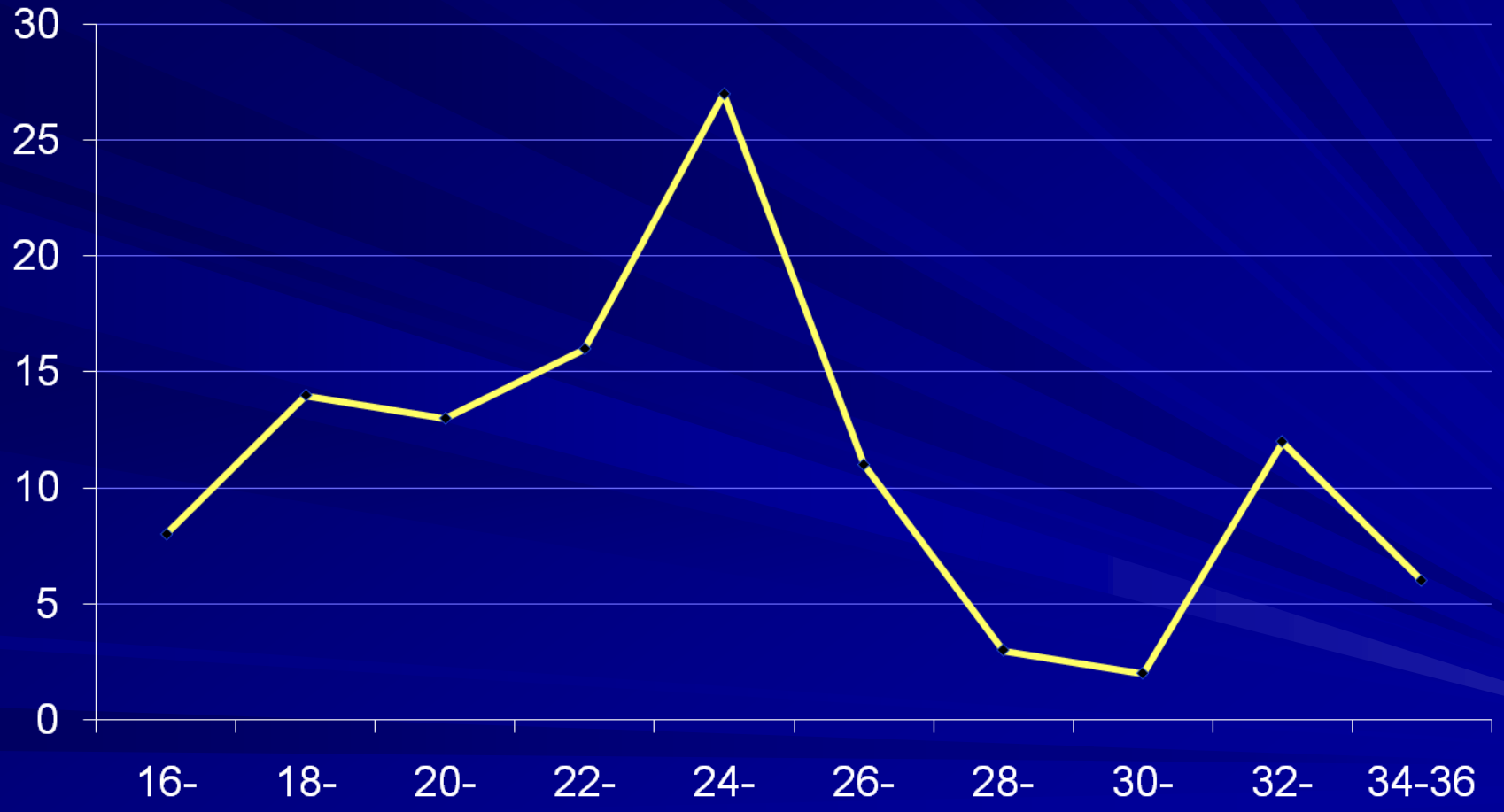
ب) المضلع التكراري أو المنحنى المنكسر: لتمثيل معلومات التوزيع التكراري نلجأ إلى مايلي: بعد رسم الاحداثيين(س وص) وبعد تقسيم محور السينات إلى فئات نضع نقاط فوق كل فئة امام محور الصادات وذلك بالاستناد إلى عدد تكرارات كل فئة، وبعد تحديد النقاط نصل بينهم بخطوط مستقيمة لنحصل على المضلع التكراري

ج) المنحنى التكراري: لرسم المنحنى التكراري الذي يعبر عن معلومات توزيع تكراري نبدأ من الخطوة الاخيرة لرسم المضلع التكراري اي بعد رسم المضلع التكراري، نعمد إلى اصلاح المنحنى المنكسر ونجعله منحنى

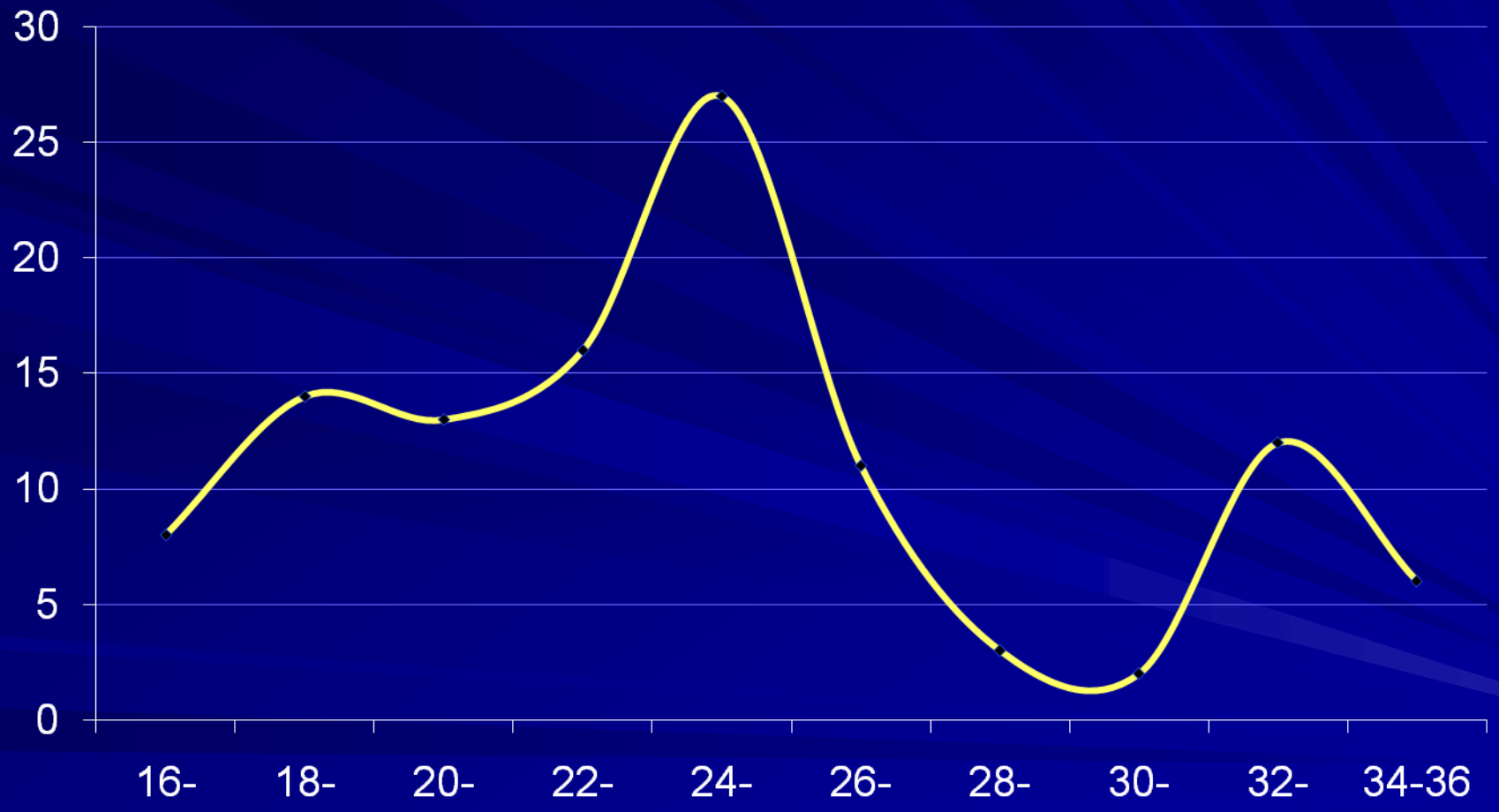
المدرج التكراري



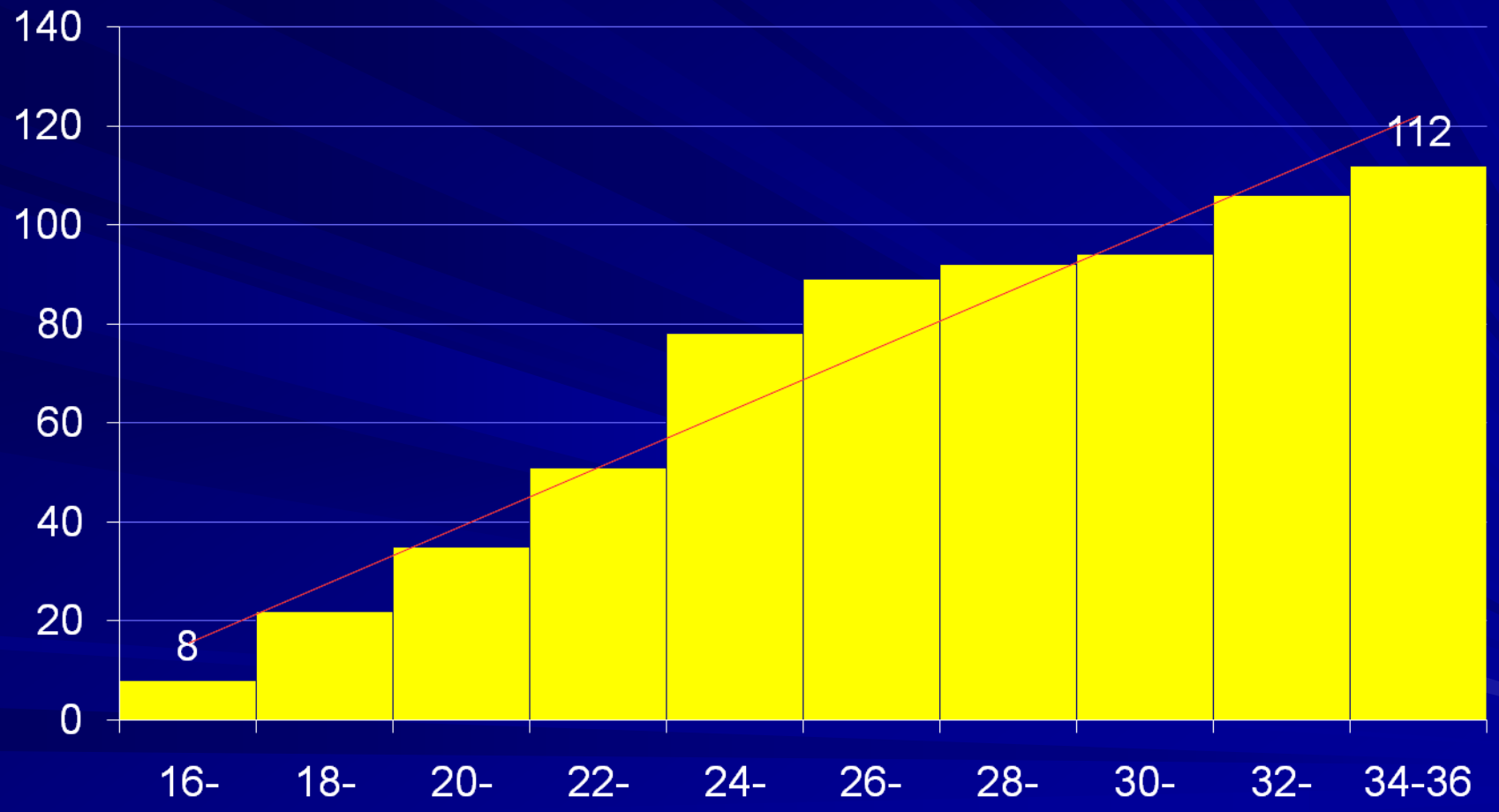
المضلع التكراري أوالمنحنى المنكسر



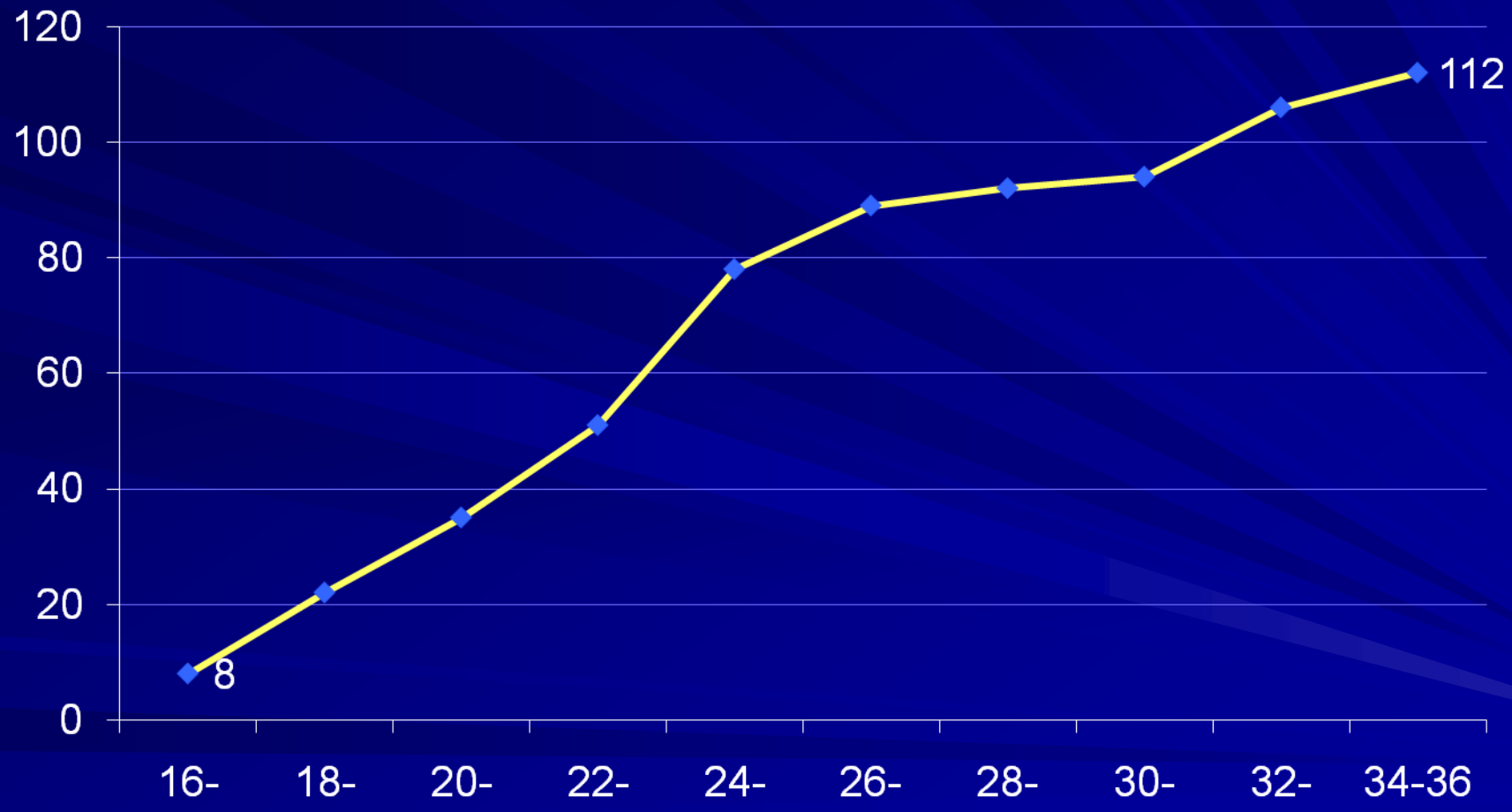
المنحنى التكراري



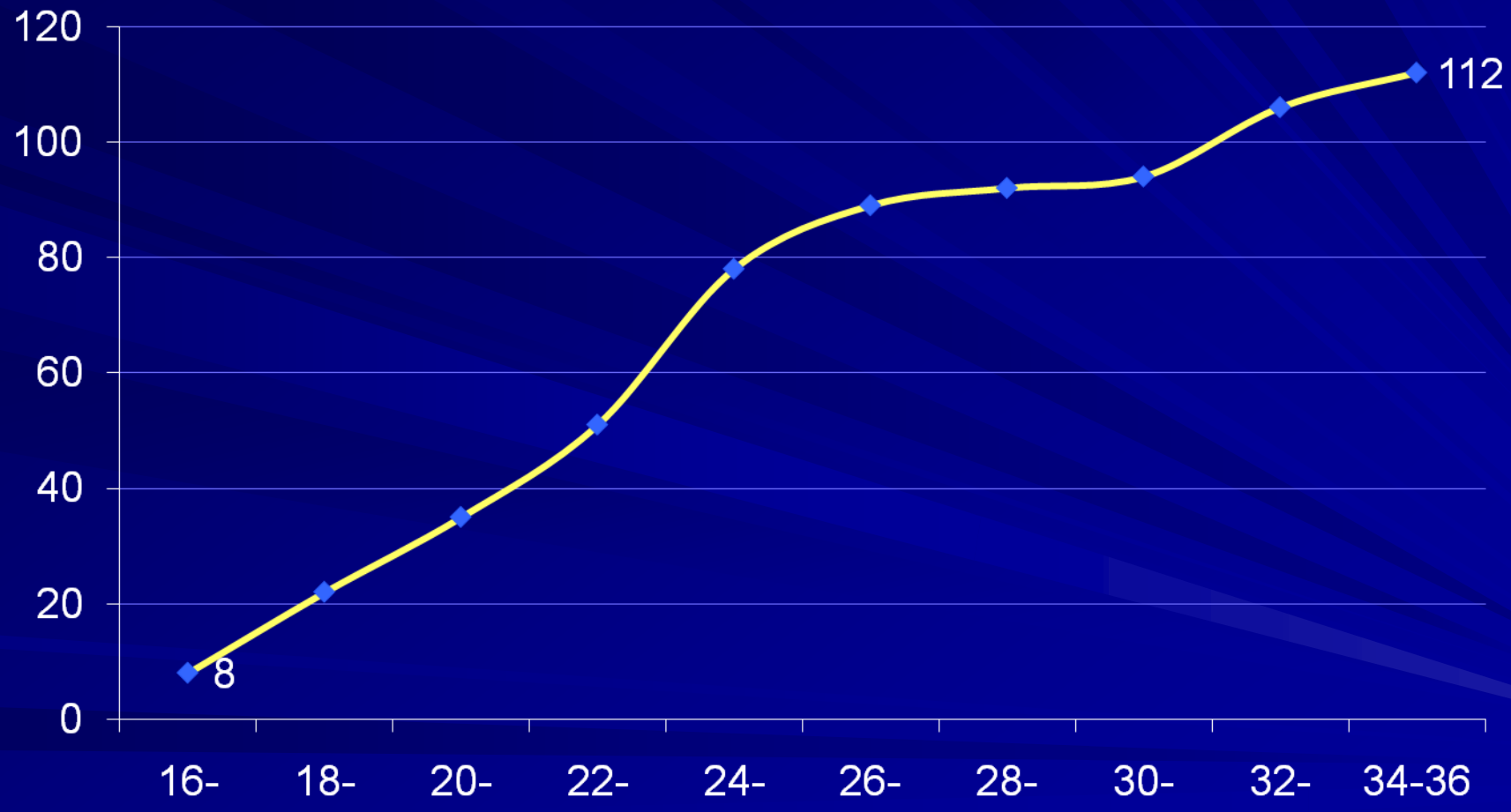
المدرج التكراري الصاعد والهابط



المضلع التكراري أو المنحنى المنكسر الصاعد والهابط



المنحنى التكراري الصاعد والهابط



مقاييس النزعة المركزية

- إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - The basics
      - Central tendency

مقاييس النزعة المركزية

لاحظ علماء الاحصاء تركز غالبية الظواهر الطبيعية والاقتصادية والاجتماعية حول قيم وسطى، لذلك استخدم العلماء هذه القيم كأدوات لتمثيل وتلخيص المعلومات الرقمية وأطلقوا على هذه القيم اسم المتوسطات.

انواع المتوسطات:

1) الوسط الحسابي
2) الوسيط
3) المنوال
4) الوسط الهندسي
5) الوسط التوافقي

لكننا هنا سنتعرض وبايجاز إلى مقياسين فقط هما الوسط الحسابي والوسيط، كتعريف وطريقة حساب وخواص وميزات وعيوب، وذلك لاعتقادنا بأنهما أكثر المتوسطات استخداماً.

الوسط الحسابي

- يعرف الوسط الحسابي لمجموعة من القيم بأنه مجموع هذه القيم مقسوماً على عددها
- كما يمكن تعريفه بأنه القيمة التي لو أُعطيت لكل مفردة من مفردات المجموعة فإن مجموع القيم الجديدة يساوي مجموع القيم الاصلية
- طرق حسابه:

أولاً: للبيانات غير المبوبة

الوسط الحسابي لبيانات غير مبوبة يساوي مجموع قيم الظاهرة مقسوماً على عددها

الوسط الحسابي

- μ المتوسط الحسابي للمجتمع

$$\mu = \sum x \div N$$

- المتوسط الحسابي من العينة

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

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

الوسط الحسابي المرجح

يطلق على المتوسط أسم الوسط الحسابي العادي تمييزاً له عن وسط حسابي آخر يطلق عليه الوسط الحسابي المرجح وهو الذي نستعمله لحساب الوسط الحسابي لعدد من القيم تتفاوت من حيث أهمية كل منها

يستخدم الوسط الحسابي المرجح في ايجاد متوسط المعدلات، كما ويستخدم في تركيب الارقام القياسية

ثانياً: الوسط الحسابي للبيانات المبوبة

خواص الوسط الحسابي - هناك خاصيتان أساسيتان للوسط الحسابي:

1) أن المجموع الجبري لانحرافات مفردات الظاهرة عن وسطها الحسابي يساوي صفراً
2) مجموع مربعات انحرافات مجموعة من القيم عن وسطها الحسابي أصغر من مجموع مربعات انحرافات هذه المجموعة عن أي قيمة أخرى

ميزات الوسط الحسابي:

- الوسط الحسابي هو أهم مقاييس النزعة المركزية وأكثرها استخداماً
- أنه مُعرف جبرياً ويمكن فهمه واستيعابه من قبل عدد كبير من القراء

عيوب الوسط الحسابي:

- أما من ناحية العيوب فأن الوسط الحسابي يتأثر بالقيم المتطرفة
- أنه يصعب استخدامه في قياس النزعة المركزية في حالة الجداول التكرارية المفتوحة

الوسيط

يعرف الوسيط لمجموعة من القيم مرتبة ترتيباً تصاعدياً أو تنازلياً بأنه القيمة التي تقسم هذه المجموعة إلى قسمين بحيث يكون عدد القيم الأقل منها يساوي عدد القيم الأكبر منها، كما يعرف الوسيط لمجموعة من القيم مرتبة ترتيباً تصاعدياً أو تنازلياً بأنه القيمة الوسطى أو متوسط القيمتين الوسطيتين.

أسلوب حسابه

يحسب الوسيط إما رياضياً أو بيانياً

ميزات الوسيط

1) معناه سهل رغم عدم شيوعه
2) حسابه سهل
3) يمكن حسابه من جداول التوزيع التكراري المفتوحة
4) قيمته محددة إذ أن 50 % من المفردات أقل منه و 50 % اكبر من قيمة الوسيط لذلك فهو متوسط مكاني وليس متوسطاً محسوباً كالوسط الحسابي
5) قيمته لاتتأثر بالقيم الشاذة
6) أن مجموع الانحرافات المطلقة عن الوسيط أقل ما يمكن أي أقل منها عن أي قيمة أخرى

عيوب الوسيط

1) غير شائع
2) من الضروري لحسابه أن ترتب البيانات تصاعدياً أو تنازلياً
3) لايمكن الاستفادة منه حسابياً
4) غير حساس للتغيرات في قيم المفردات الداخلة في التوزيع فقد نغير في قيم هذه المفردات دون أن يتغير الوسيط

مقاييس التشتت أو التنوع

- إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - The basics
      - Variability

مقاييس التشتت أو التنوع

- يقصد بالتشتت أو التباين لأي مجتمع أوعينة بأنه التباعد أو الاختلاف بين مفردات هذا المجتمع أوهذه العينة
- وتكون قيمة التشتت أو التباين مساوية للصفر إذا لم يكن هناك اختلاف بين المفردات، أي إذا كانت جميع المفردات متساوية في قيمتها
- ويكون التباين كبيراً كلما زادت الاختلافات بين المفردات وبعدت في قيمتها عن متوسطها الحسابي
- وعلى هذا يعتبر تشتت القيم أو تباينها كمقياس لتركز المفردات حول المتوسط أو مقياساً لتجانسها

مقاييس التشتت

مقاييس التشتت كثيرة وتختلف من حيث دقتها وطرق حسابها وأهمها:

1) المدى
2) الانحراف المتوسط
3) التباين ← الانحراف المعياري
4) الخطأ المعياري
5) معامل الاختلاف

لكننا سنتناول هنا اكثرها انتشارا واستخداما ألا وهو الانحراف المعياري

الانحراف المعياري

هو مقياس للتشتت أو التباين حول الوسط الحسابي وهو عبارة عن الجذر التربيعي لمتوسط مربعات انحرافات مجموعة من القيم عن وسطها الحسابي أي هو عبارة عن الجذر التربيعي للتباين

والانحراف المعياري أهم مقاييس التشتت وأكثرها استخداماً ويرمز له σ إذا حسب للمجتمع بالرمز s إذا حسب للعينة

معادلات لتقدير التباين

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

$$S_x = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}$$

معادلات الانحراف المعياري

$$s_x = \sqrt{s^2}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

$$S_x = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}}$$

معامل الاختلاف

إذا كان مقياس التشتت المطلق هو الانحراف المعياري ومقياس التوسط هو الوسط الحسابي فان مقياس

التشتت النسبي يسمى معامل الاختلاف ويعرف كمايلي:-

$$\text{معامل الاختلاف} = \frac{\text{الانحراف المعياري}}{\text{الوسط الحسابي}} \times 100$$

يستعمل لمقارنة التشتت بين مجموعتين مفردات كل منهما تقاس بوحدات مختلفة

الارتباط الخطي البسيط

- إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - Regression
      - Linear correlation

الارتباط

سوف نناقش في الجزء اللاحق كيفية استخدام تحليل الانحدار لتحديد العلاقة بين متغيرين بصيغة معادلة رياضية ومن ثم استخدام هذه المعادلة (معادلة التنبؤ) للتنبؤ بالمتغير التابع عند معرفة المتغير المستقل.

لكننا في بعض المسائل نكون معنيين فقط بمعرفة العلاقة بين المتغيرات. مثل درجة العلاقة بين الدخل القومي والواردات، درجة العلاقة بين المبيعات ونفقات الاعلان...الخ.

ويمكن قياس مدى الارتباط بين صفتين عن طريق حساب معامل الارتباط للعينة بالرمز(r) لانه عبارة عن قيمة تقديرية للمجتمع م تنحصر قيمة معامل الارتباط بين + 1 و- 1 وهو قيمة مطلقة لا يعبر عنها بوحدات القياس.

إذا كان الارتباط بين الصفتين بالزيادة بمعنى أن احدى الصفتين تزيد بزيادة الاخرى سمي ارتباطا موجبا وقيمته بين (صفر و+ 1)

واذا كان الارتباط بين الصفتين بالنقص بحيث تتناقص قيم احدى الصفتين بزيادة الاخرى سمي ارتباطا سالبا وقيمته بين (صفر و- 1)

أما إذا لم تكن هناك علاقة بين الصفتين فأن قيمة معامل الارتباط = صفر

ولذلك فأن من الاهمية بمكان ملاحظة اشارة معامل الارتباط وليست قيمته فقط

أنواع الارتباط:

أولا- يقسم الارتباط من حيث قوته إلى:

1) الارتباط الكامل: وهذا يعني أن تغير احد الظواهر أو بعضها يتوقف كليا على تغير الاخرى

ومثال ذلك العلاقة بين مساحة الدائرة ونصف قطرها، ويمكن أن تسمى العلاقة بين هذه المتغيرات علاقة دالية

2) الارتباط الجزئي: وهذا النوع يعني أن هناك بعض الارتباط بين المتغيرات المختلفة ولكن هذه العلاقة ليست بقوة العلاقة الدالية ومن الامثلة على ذلك: العلاقة بين مستوى الدخل في بلد ما والانفاق على سلعة معينة، العلاقة بين الانفاق على سلعة معينة ومستوى الدخل وسعر هذه السلعة وسعر السلع البديلة والتقاليد والعادات، العلاقة بين الطول والوزن، العلاقة بين الحقن بلقاح خاص واكتساب مناعة ضد مرض معين، العلاقة بين البطالة والاجرام.

وفرض العلاقة بين الظواهر المختلفة ينشأ من مجرد الاستنتاج المنطقي أو مشاهدة مجموعة من الظواهر وملاحظة ان تغير احدهما أو مجموعة منها مرتبطة بتغير الاخرى.

ثانيا- يقسم الارتباط من حيث عدد المتغيرات التي تؤخذ في الاعتبار إلى:

1) الارتباط البسيط: وفي هذا النوع ندرس العلاقة بين متغيرين فقط.
2) الارتباط المتعدد: وفي هذا النوع ندرس العلاقة بين اكثر من متغيرين.

ثالثاً- أما من حيث شكل العلاقة الرياضية التي تربط بين المتغيرات المختلفة فأن الارتباط يقسم إلى:

1) ارتباط خطي
2) ارتباط غير خطي

معادلات لتقدير معامل الارتباط الخطي البسيط

$$r_{yx} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (x_i - \bar{x})^2}}$$

$$r_{yx} = \frac{\sum_{i=1}^{n} (yx)}{\sqrt{(\sum y^2)(\sum x^2)}}$$

الانحدار الخطي البسيط

- إذهب إلى
- StatTrek.com
  - AP Statistics - Exploring Data
    - Regression
      - Linear regression

الانحدار

- يعتمد تحليل الانحدار على العلاقة بين متغيرين أو اكثر. والتحليل هنا يقوم على أساس وجود متغير تابع وآخر مستقل (متبوع). فبمجرد تحديد العلاقة الرياضية بين المتغيرين يسهل تحديد المتغير التابع بمعرفة بيانات المتغير المستقل (المتبوع).
- فاذا كانت الواردات مثلا تتأثر بالدخل القومي فبتحديد هذه العلاقة كميا يمكن التنبوء بالواردات بمجرد معرفة الدخل القومي المتوقع.
- ومن الناحية الرياضية إذا اعتمد المتغير التابع Y في قيمته على مقدار التغير في قيمة المتغير المستقبل X فانه يعبرعن Y بانها دالة في X وهو مايسمى بالانحدار.
- أما معامل الانحدار فهو المؤشر الذي يبين لنا مدى التغير الذي يحصل في متغير ما (المتغير التابع) نتيجة تغير وحدة واحدة من المتغير الاخر(المتغير المستقل).

الانحدار الخطي

هب اننا معنيين بدراسة العلاقة بين قيم متغيرين مختلفين كالعلاقة بين الدخل والواردات أو الزمن والصادرات أو الاعلان والمبيعات...الخ فأن الخطوة الاولى لدراسة هذه العلاقة هي جمع البيانات ثم *تمثيلها بيانيا* *X,Y* هي:

(xn, yn)..... , (x3, y3), (x2, y2), (x1, y1)

ومثلنا Xعلى المحور الافقي (المتغير المستقل) وY على المحور العمودي (المتغيرالتابع)

ومن الجدير بالذكر انه ليس من الضروري أن تقع جميع النقط على خط الانتشار أو تكون ملاصقة له اذ يمكن أن تنحرف بعض النقط عن هذا الخط لسبب ما وتسمى في هذه الحالة نقط شاذة أو متطرفة ولكي يكون الخط الناتج (مستقيما اوغيرمستقيم) ممثلا للبيانات المعطاة فانه يجب أن يمر بعدد كبير من هذه النقط ويتوسط الباقي احسن توسط. اي يجب أن يكون مجموع مربعات انحرافات النقط في الشكل الانتشاري عن نظيراتها على خط الانتشار(خط الانحدار) اقل مايمكن.

وللوصول إلى افضل خط يمثل العلاقة بين المتغيرين نتبع طريقة المربعات الصغرى ومنها يمكن حساب افضل خط يربط هذه النقط باقل انحراف ممكن وفي الوقت نفسه يمكن التعبير عنه بمعادلة يحسب منها، وهذا الخط يطلق عليه خط الانحدار

وبمعنى اخر فان استعمال طريقة المربعات الصغرى تساعد الباحث على توفيق افضل خط لمجموعة من البيانات نتيجة لان مجموع مربعات انحرافات النقط عن هذا الخط يكون اقل ما يمكن ومن هذا الخط

يمكن استخراج كمية احصائية معينة يطلق عليها معامل الانحدار(وقد سبق تعريفه) ويرمز لمعامل الانحدار للعينة (معامل انحدار y على x) بالرمز $b_{y.x}$ كما يرمز لمعامل انحدار المجتمع بالرمزB

ومعامل الانحدار هوعبارة عن ميل خط الانحدار.او هو مقدار التغير في المتغير(سواء بالزيادة اوالنقصان) نتيجة لزيادة وحدة واحدة من وحدات المتغير المستقبل.

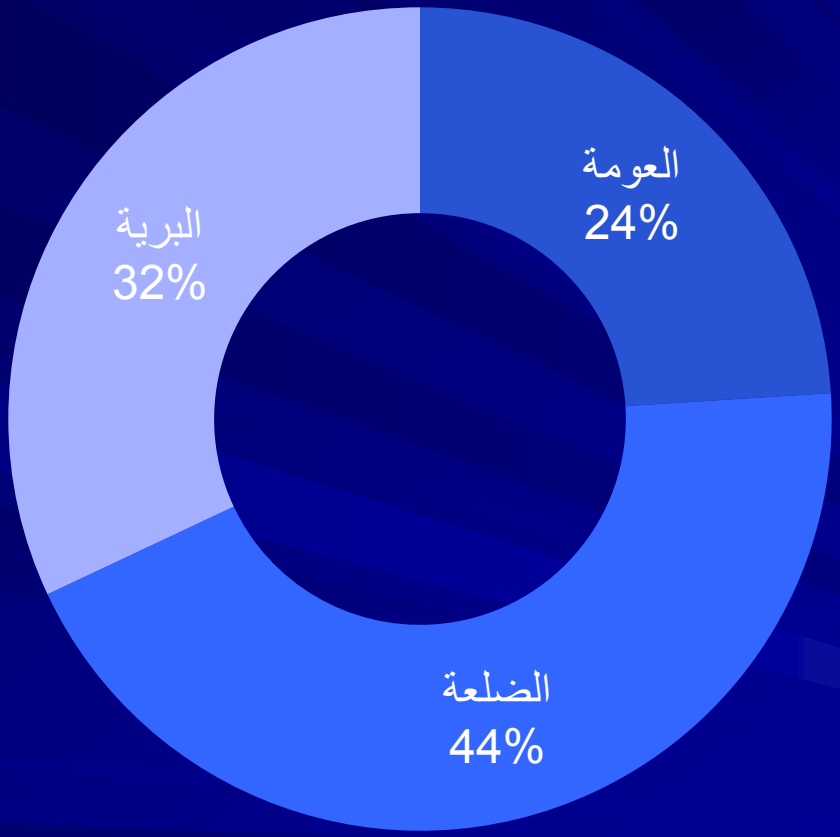
معادلات لتقدير معامل الانحدار الخطي البسيط

$$b_{yx} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

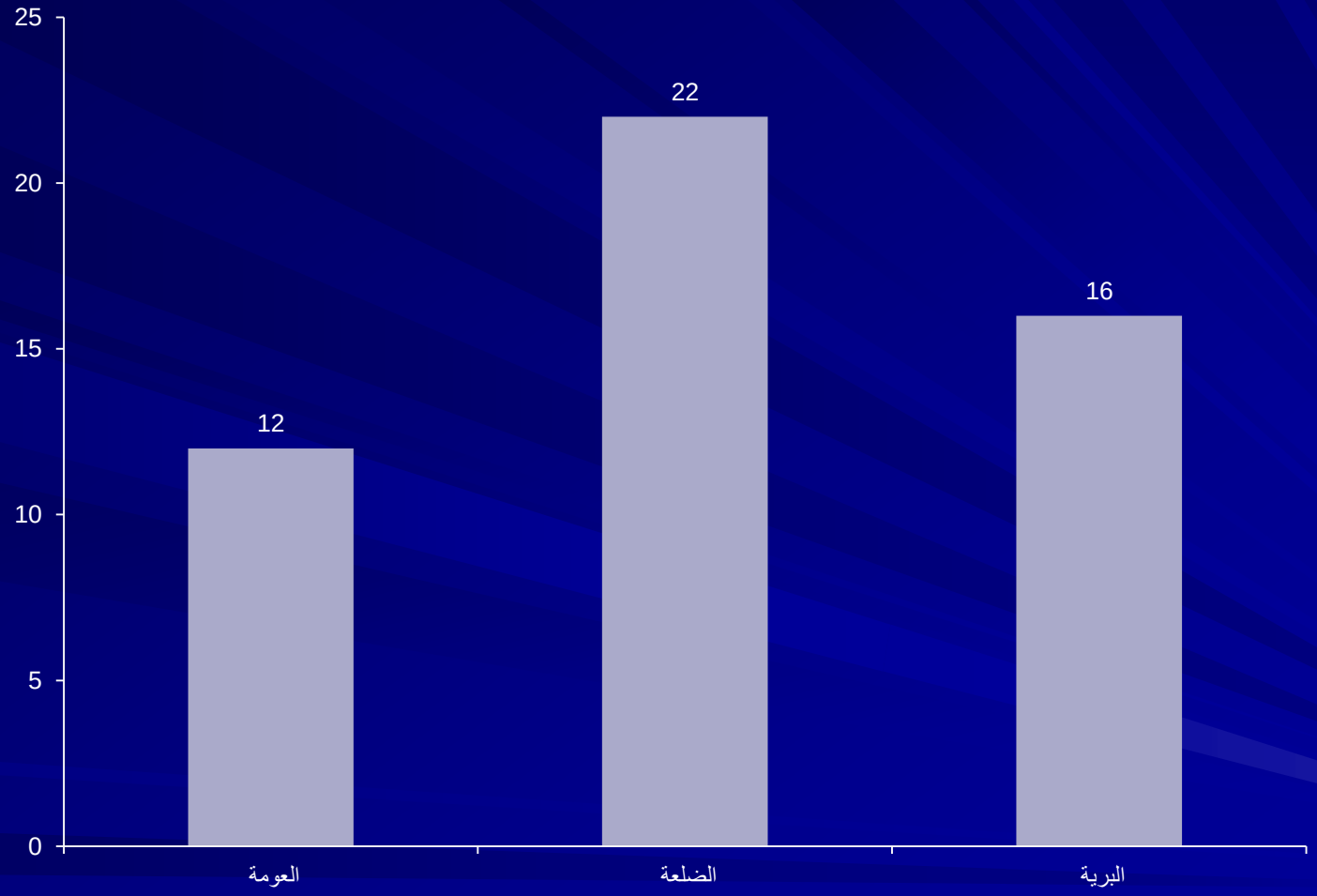
$$b_{yx} = \frac{\sum yx - \frac{(\sum y)(\sum x)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$a_{yx} = \bar{y} - b_{yx}(\bar{x})$$

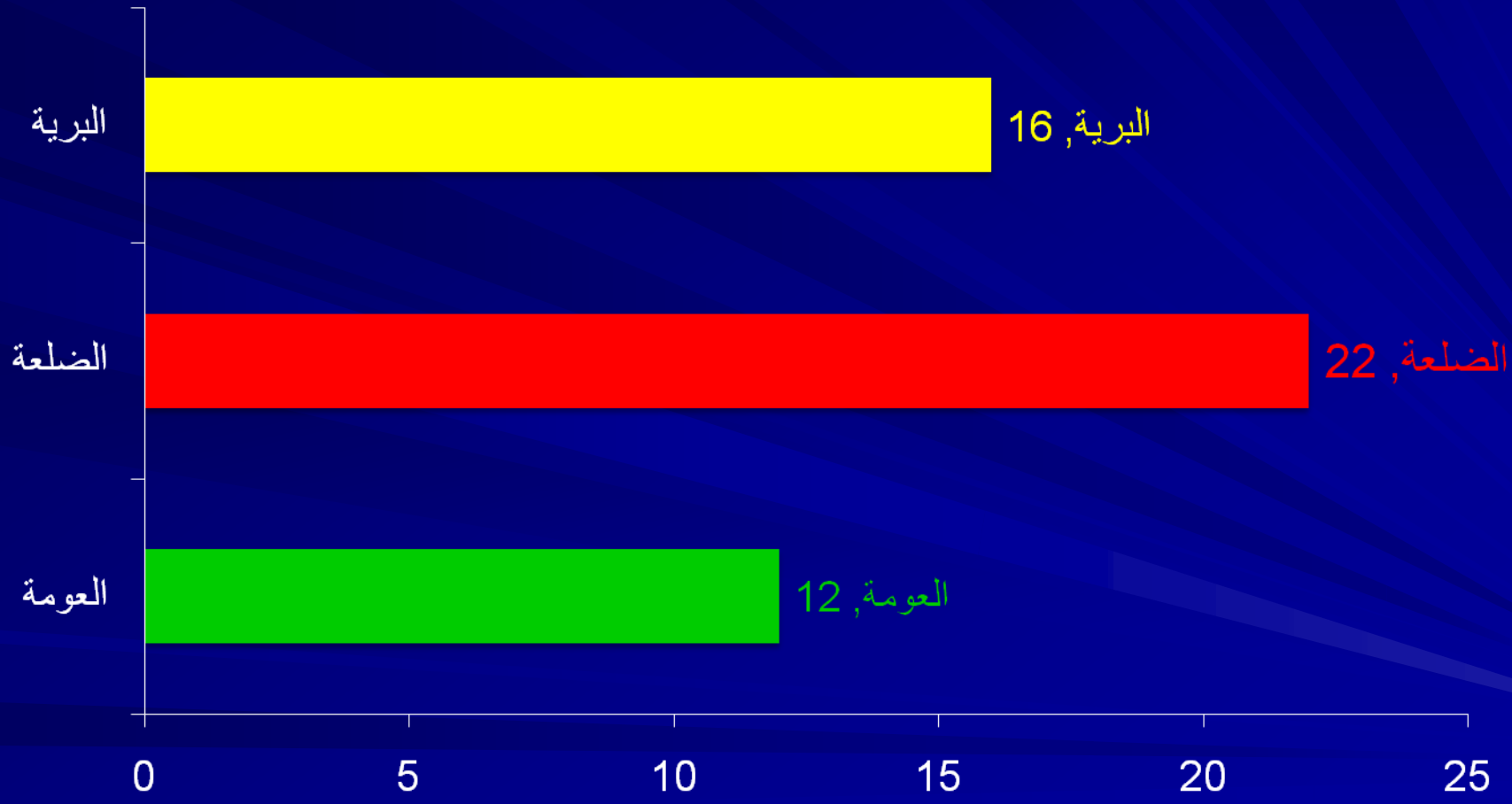
تحليل السلاسل الزمنية



شكل رقم (1) نوع الاسماك



شكل رقم (1) الاسماك حسب النوع



المصدر : وزارة الزراعة – مسح انتاج الاسماك