Beginners Guide to Statistics in Data Science

Introduction

ال statistics هي احد تخصصات الرياضيات متفق عليها عالميا بشرط أساسي لفهم أوسع و اعمق لل machine learning

مع ان ال statistics مجال كبير و فيه نظريات و نتائج كتيرة إلا ان الادوات و الملاحظات اللي بتتاخد منه مطلوبة لممارسين ال machine learning و دا مع وجود اساس ثابت لأصل الإحصاء

لو هنتكلم عن الأدوات الإحصائية اللي هنستخدمها في الممارسة العملية هيكون افضل لينا نقسم مجال ال statistics لجزأين من الاساليب جزء وصفي لتلخيص البيانات و جزء استنتاجي لاستخلاص النتائج من عينات البيانات:

- **Descriptive statistics**

و دي بتشير لطرق تلخيص الملاحظات الأولية في معلومات ممكن نفهمها و نشاركها و في منها نوعين :

- ال uni-variate descriptive statistics :

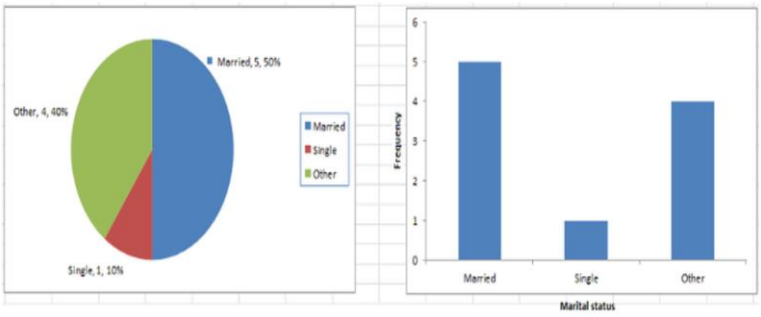
بيتضمن الطرق المختلفة اللي بنوصف بيها الأنماط الموجودة في البيانات احادية التباين

ال central tendency زي :

- Mean
- Mode
- Median

ال dispersion زي :

- Range
- Variance
- Maximum
- Minimum
- Quartiles
- Standard deviation



The various plots used to visualize uni-variate data typically are Bar Charts, Histograms, Pie Charts. etc.

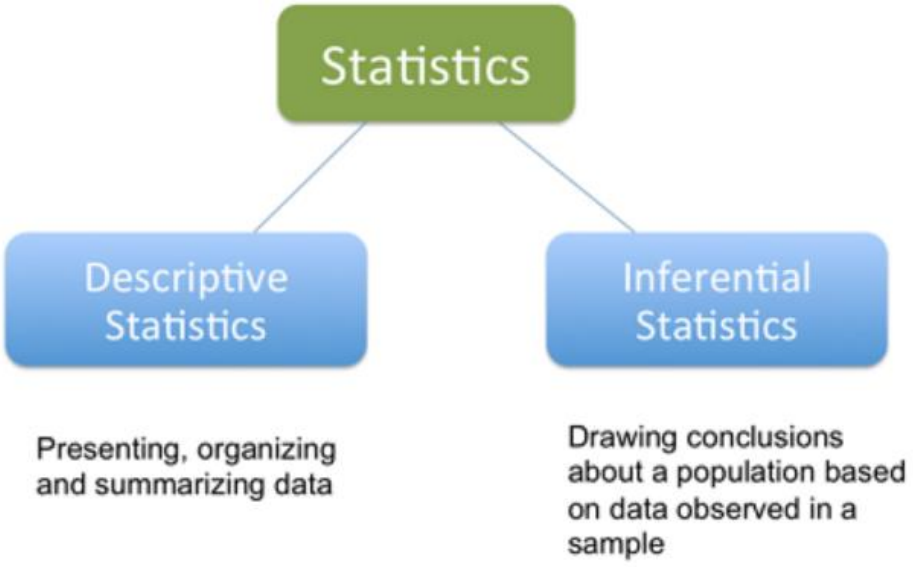
- ال bi-variate descriptive statistics :

بيتضمن تحليل ال Bi variate تحليل متغيرين لغرض تحديد العلاقة التجريبية بينهما

من ال plots المستخدمة عشان نعملها visualisation ال scatter plot و ال box plot

- **Inferential statistics**

و دي بتشير للطرق اللي بتساعدنا في تحديد خصائص المجال او السكان من مجموعة اصغر من الملاحظات بنسميها عينة



Statistical concepts

- Elements

ال entities اللي بنجمع معلومات ليها بنسميها elements و ممكن نسميها cases او subjects

- Variables

عبارة عن attribute بيصف شخص او مكان او شئ او فكرة بتاخد قيم مختلفة باختلاف ال entities

امثلة عليها : الحالة الاجتماعية و الدخل و السنة و ليها نوعين اما فئوية او رقمية

- Qualitative

عبارة عن متغير نوعي بيمكّن العناصر من تصنيفها او جعلها فئوية حسب بعض الخصائص زي الحالة الاجتماعية و الرتبة و الرهن العقاري

- Quantitative

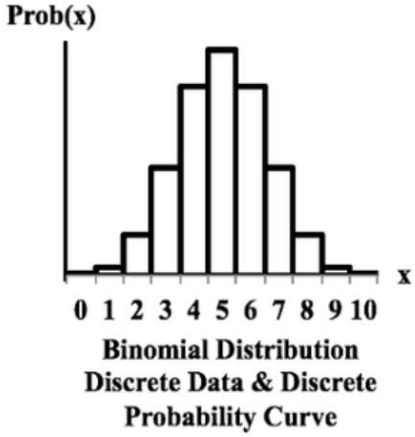
عبارة عن متغير كمي و ليه قيم رقمية و بيسمح بإجراء العمليات الحسابية بشكل هادف زي الدخل و السنة

- Discrete variable

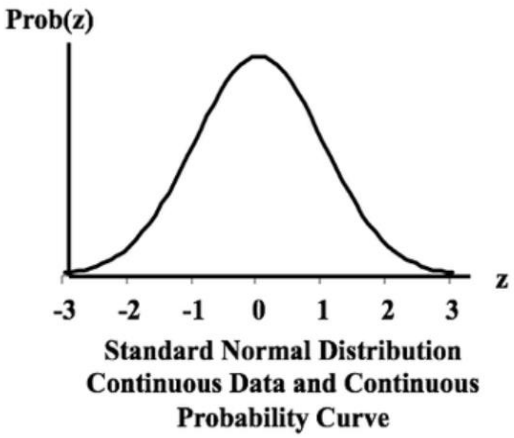
المتغير العددي اللي ممكن ياخد عدد محدد او عدد قابل للعد من القيم هو discrete variable بحيث ممكن نرسم كل قيمة ك نقطة منفصلة و في مسافة بين كل نقطة زي السنة مثلا

- continuous variable

المتغير العددي اللي ممكن ياخد عدد غير محدود من القيم هو continuous variable و القيمة المحتمل بتمثل فاصل على خط الاعداد و مفيش مسافة بين النقاط زي الدخل مثلا



Binomial Distribution
Discrete Data & Discrete Probability Curve



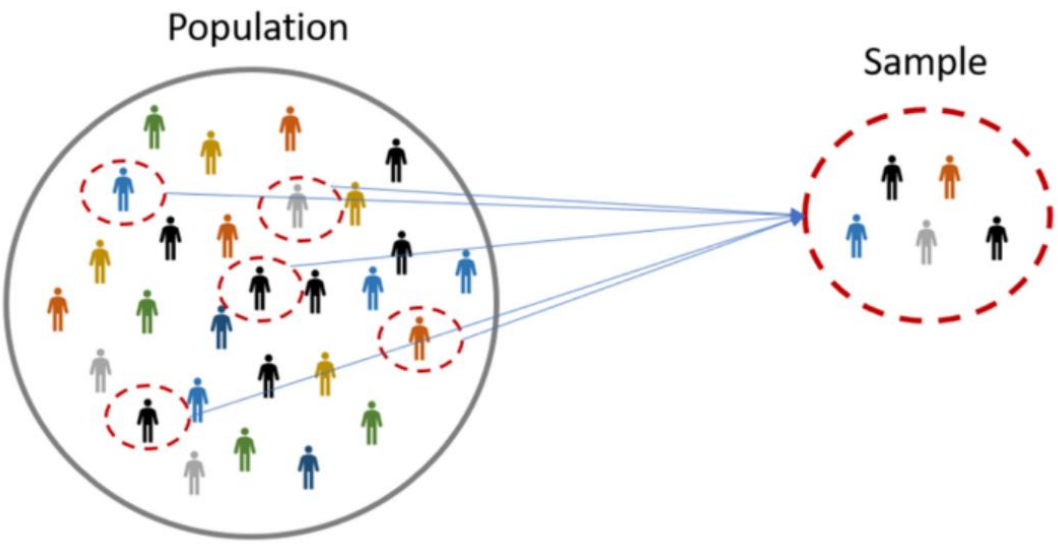
Standard Normal Distribution
Continuous Data and Continuous Probability Curve

- population

هو مجموعة كل عناصر الاهتمام لبعض المشاكل
ال parameter هي خاصية مميزة للسكان

- Sample

بتتكون من مجموعة فرعية من ال population
خاصية العينة بنسميها إحصائية



- Random sample

لما ناخد عينة بيكون لكل عنصر فيها فرصة متساوية في الاختيار

How to measure central tendency ?

للإشارة إلى مكان تحديد الجزء المركزي من البيانات على خط الارقام بنستخدم :

- Mean

هو متوسط مجموعة البيانات تجمع القيم و اقسمها على عدد القيم

متوسط العينة هو المتوسط الحسابي للعينة و بنرمز له ب x-bar

متوسط المجتمع هو متوسط عدد السكان

- Median

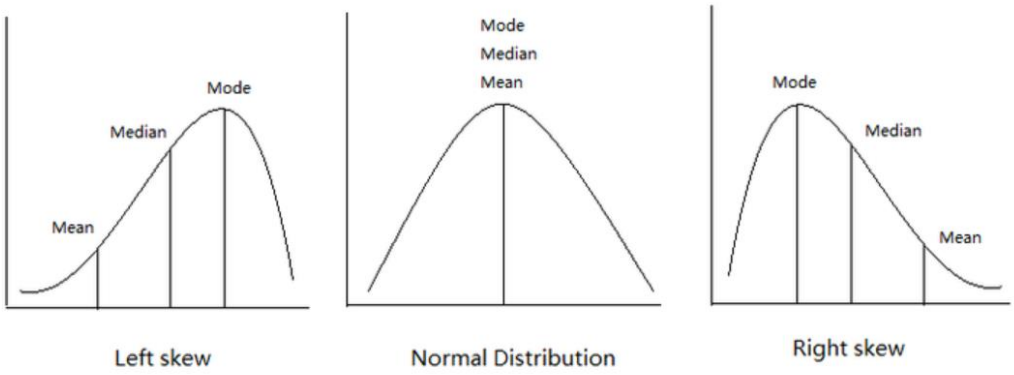
الوسيط هو قيمة البيانات المتوسطة لما يكون فيه عدد فردي من قيم البيانات و بيتم فرز البيانات بترتيب تصاعدي

لو فيه رقم زوجي ف الوسيط هو متوسط قيمتى البيانات الوسطيتين لما يتم فرز البيانات بترتيب تصاعدي

- Mode

هو القيمة الأكثر شيوعا في مجموعة البيانات

المتغيرات الكمية و الفئوية عادي يكون ليها mode انما المتغيرات الكمية بس اللي ليها mean و median



- Mid-range

هو متوسط الحد الأقصى و الحد الأدنى في مجموعة البيانات

How to measure variability?

لوصف مقدار التباين او الانتشار في البيانات نستخدم :

- Range

الفرق بين الحد الأقصى و الحد الأدنى للقيم في مجموعة من القيم

ال range بيعكس الفرق بين الملاحظة الأكبر و الأصغر لكنه بيفشل يشير إلى كيفية مركزية البيانات

- Variance

في السكان بيتم تعريف ال variance على إنه متوسط الانحراف التربيعي عن المتوسط

ال variance الاكبر معناه ان البيانات اكثر انتشارا

ال sample variance هو تقريبا متوسط الانحرافات التربيعية بحيث بيتم استبدال ال N ب n-1 و الاختلاف دا بسبب استخدام متوسط العينة كتقريب لمتوسط المحتوى الحقيقي

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- Standard deviation

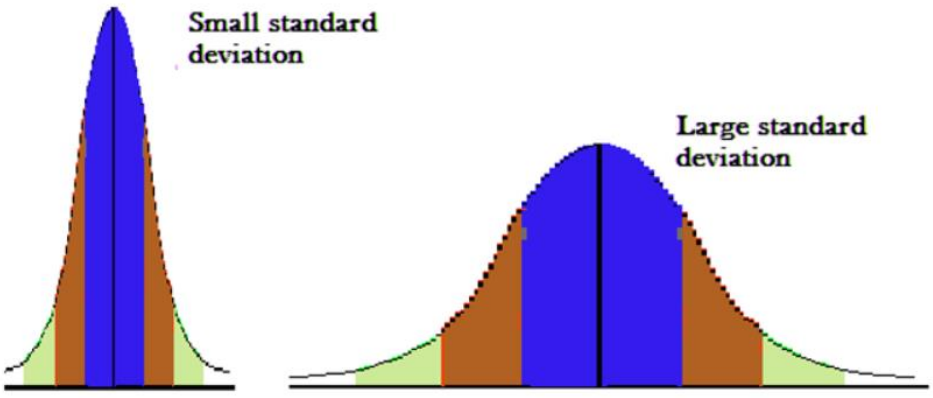
الانحراف المعياري لمجموعة من الارقام بيعرفنا إلى أي مدى الارقام الفردية بتميل للإختلاف عن المتوسط

الانحراف المعياري للعينة هو الجذر التربيعي ل variance العينة

الانحراف المعياري لل population هو الجذر التربيعي ل variance ال population

كلما كان الانحراف المعياري اصغر كلما كانت نقاط البيانات اقرب الى المتوسط

كلما كانت نقاط البيانات بعيدة عن الوسط كلما زاد الانحراف المعياري



How to measure position ?

لفهم اين تقع قيمة معينة في عينة او توزيع نستخدم :

- Deciles

بيقسم البيانات إلى 10 اجزاء متساوية و كل جزء بيمثل 10/1 من البيانات

- Percentile

ال pth percentile لمجموعة البيانات هي قيمة البيانات بحيث تكون النسبة المئوية للقيم في مجموعة البيانات عند هذه القيمة أو أقل منها

- Interquartile range (IQR)

الربع الاول (Q1) هو النسبة المئوية الخامسة و العشرون لمجموعة البيانات

الربع الثاني (Q2) هو النسبة المئوية الخمسين يعني الوسيط

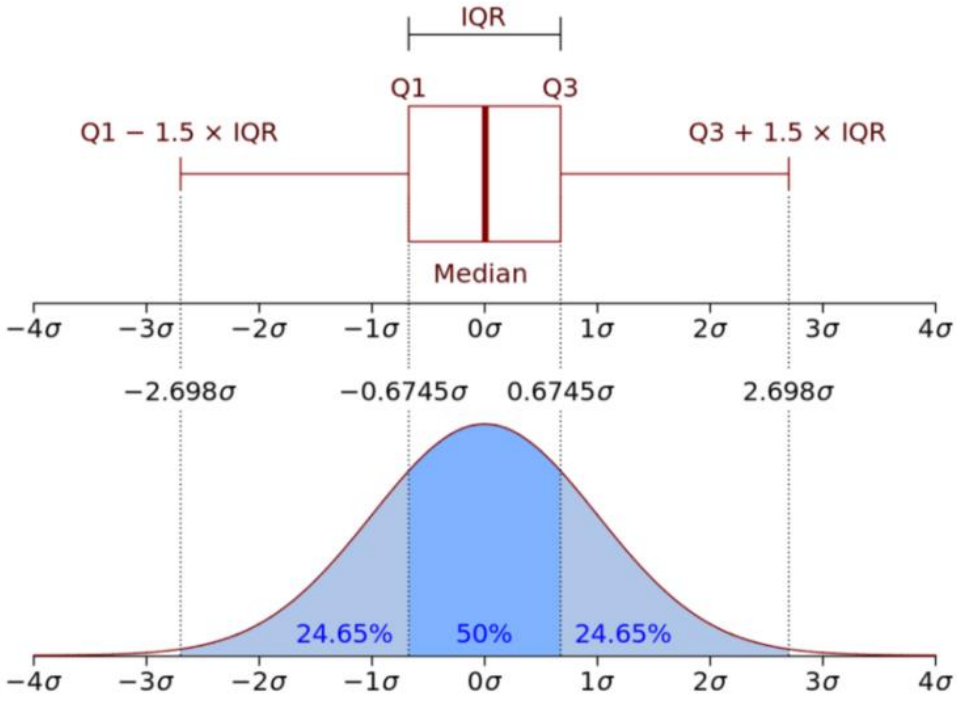
الربع الثالث (Q3) هو النسبة المئوية الخامسة و السبعون

ال IQR بيقيس الفرق بين Q3 و Q1

IQR = Q3 - Q1

قيمة البيانات x بتبقى outlier لو :

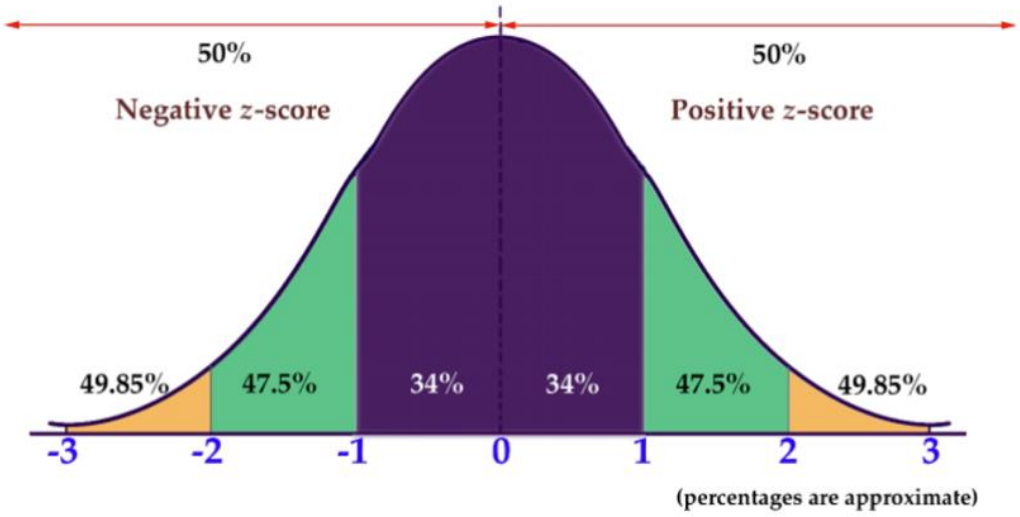
X <= Q1 - 1.5(IQR) or X >= Q3 + 1.5(IQR)



- Standard score or Z-score

ال z-score بتمثل لقيمة بيانات معينة عدد الانحرافات المعيارية التي تقع قيمة البيانات اعلى و اسفل المتوسط

لو ال Z موجبة ف دا معناه ان القيمة اعلى من المتوسط



- Scatter plots

ابسط طريقة لتصور العلاقة بين متغيرين كميين x و y

بالنسبة لمتغيرين مستمرين ف ال Scatter plots هو رسم بياني شائع

- Correlation

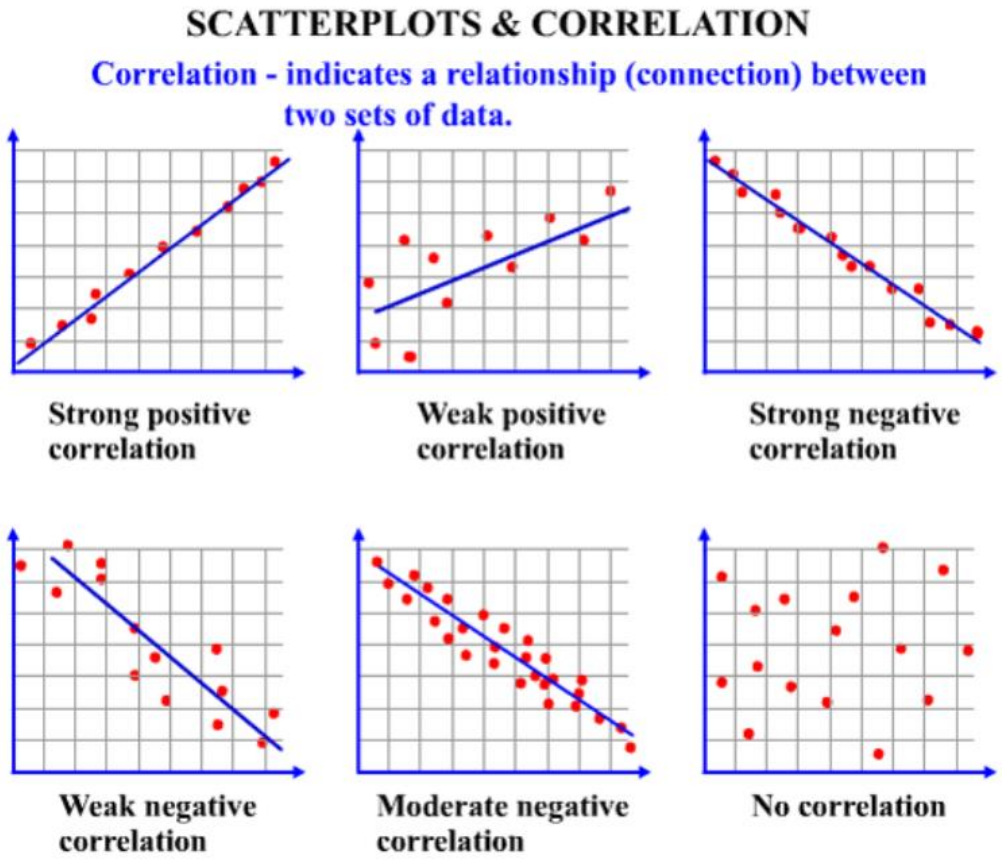
هو احصاء بتهدف لتحديد قوة العلاقة بين متغيرين

معامل ال correlation بيقيس قوة و اتجاه العلاقة الخطية بين متغيرين كميين

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)\, s_x s_y}$$

لو r موجبة ف هنقول ان في ارتباط ايجابي بين x و y و الزيادة في x ترتبط بزيادة في y

لو r سالبة ف هنقول ان في ارتباط سلبي بين x و y و الزيادة في x بترتبط بانخفاض في y

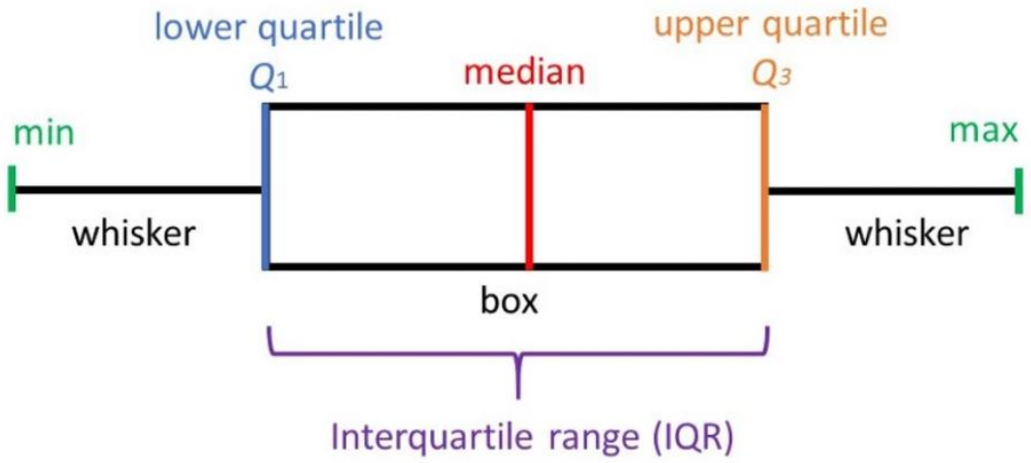


- Box plots

بيستخدم لتصوير توزيع القيم لما يكون احد المتغيرات فئوي و الاخر مستمر

بيقوم بتقسم قيم البيانات إلى اربعة اجزاء بنسميها quartiles

وسيط كل نصف بيؤدي الى تقسيم قيم البيانات إلى اربعة أجزاء



Population and sampling Methods

ال population بيحتوي على جميع نقاط البيانات من مجموعة من البيانات بينما ال Sample بتتكون من عدد قليل من الملاحظات المختارة من المجتمع

اختيار العينة من المجتمع لازم تحتوي على جميع الخصائص اللي بيمتلكها المجتمع

بننشئ عينات باستخدام ال Sampling methods و دي ممكن تكون probability based او non probability based

-probability sampling

تقنية لأخد العينات بيتم فيها جمع عينات من عدد كبير من السكان باستخدام probability theory و في 3 انواع لأخد العينات :

- Random sampling

في الطريقة دي كل فرد من المجتمع بيبقى ليه فرصة متساوية لأختياره في العينة

- Systematic sampling

في الطريقة دي بيتم اختيار nth record من السكان ليكون جزء من العينة

- Stratified sampling

في الطريقة دي بيتم استخدام طبقة لتكوين عينات من عدد كبير من السكان

الطبقة هي مجموعة فرعية من السكان تشترك في خاصية واحدة على الأقل

بعد كدا بيتم استخدام ال random sampling لاختيار عدد كافي من الموضوعات من كل طبقة

Why do we need inferential statistics?

على عكس ال descriptive statistics بدل ما نوصل لجميع السكان غالبا بيكون عندنا عدد محدود من البيانات

ال inferential statistics بتدخل في حيز التنفيذ يعني مثلا لو مهتمين نلاقي متوسط درجات امتحان لمدرسة بأكملها ف دا غير معقول لأن مش عملى اننا نلاقي كل البيانات اللي هنحتاجها ف الحل اننا نقيس عينة اصغر من الطلاب العينة دي هتصف المجموعة الكاملة لجميع طلاب المدرسة

ببساطة ال inferential statistics بتقوم بعمل تنبؤات حول مجموعة سكانية بناء على عينة من البيانات اللي اخدناها من المجتمع دا

تقنيات ال inferential statistics بتتم على خطوتين :

- بناخد بعض العينات و بنحاول نلاقي عينة بتمثل جميع السكان بدقة

- بعد كدا بنختبر العينة و بنستخدمها لرسم تعميمات حول المجتمع ككل

ال inferential statistics ليها هدفين :

- Estimating parameters

بناخد احصائية من البيانات اللي جمعناها زي الانحراف المعياري و بنستخدمها لتحديد parameter عام زي الانحراف المعياري للمجتمع الكامل

- Hypothesis testing

مفيد جدا لما نبقى عاوزين نجمع بيانات عن شئ منقدرش نديه غير لمجموعة سكانية محصورة جدا يعني مثلا لو عاوزين نعرف الدواء هيجيب نتيجة مع جميع المرضى ف ممكن نستخدم البيانات اللي جمعناها للتنبؤ ب دا

Statistical terminologies

- Statistic

مقياس واحد لبعض سمات العينة زي الوسيط او المتوسط

- Population statistic

احصاء السكان بالكامل في سياق زي مثلا رواتب جميع سكان علماء البيانات في مصر

- Sample statistic

احصائية مجموعة مأخوذة من السكان

- Standard deviation

مقدار التباين في بيانات السكان

- Standard Error

هو مقدار التباين في بيانات العينة و يرتبط بالانحراف المعياري

Probability

يشير احتمال وقوع حدث الي احتمال وقوع الحدث من اهم مصطلحاتها :

- Random Experiment

التجربة العشوائية او التجربة الإحصائية هي تجربة بتكون فيها جميع النتايج المحتملة للتجارب معروفة فعلا

يمكن تكرار التجربة عدة مرات في ظل ظروف متطابقة او متشابهة

- Sample space

مساحة العينة لتجربة عشوائية هى مجموعة جميع النتائج المحتملة لتجربة عشوائية

- Event

مجموعة فرعية من مساحة العينة تسمى حدثا

- Trail

تشير التجربة إلى نوع خاص من التجارب بيكون فيها نوعين من النتايج المحتملة النجاح او الفشل مع تفاوت احتمالية النجاح

- Random variable

المتغير اللي قيمته بتخضع للتغيرات بسبب العشوائية بنسميه متغير عشوائي ليه نوعين اما متغير منفصل او متغير مستمر

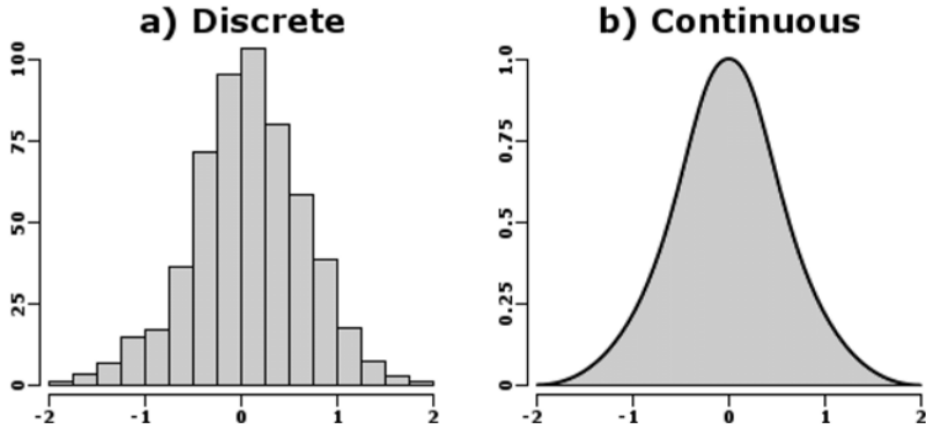
Conditional probability

الاحتمال الشرطي هو احتمال حدوث حدث معين A بالنظر الى حالة معينة حدثت فعلا B ف الاحتمال الشرطي P بنعرفه من خلال :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Probability distribution and distribution function

ال mathematical function اللي بتصف عشوائية متغير عشوائي بنسميها التوزيع الاحتمالي يعني تصوير لجميع النتايج المحتملة لمتغير عشوائي و الاحتمالات المرتبطة بها



Sampling distribution

التوزيع الاحتمالي للاحصاءات لعدد كبير من العينات المختارة من السكان بنسميه توزيع العينات

لما يزيد حجم العينة متوسط العينة بيبقى اكثر توزيعا بشكل طبيعي حول متوسط السكان و بيقل تنوع العينة كل ما زاد حجم العينة

Normal distribution

التوزيع الطبيعي هو توزيع احتمالي مستمر موصوف بال normal equation

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \quad e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

منحنى التوزيع الطبيعي متماثل على جانبي الوسط يعني الجانب الأيمن من المركز هو صورة معكوسة للجانب الأيسر

المنطقة اللي تقع اسفل منحنى التوزيع الطبيعي تمثل الاحتمالية و المساحة الإجمالية تحت المنحنى بتساوي 1

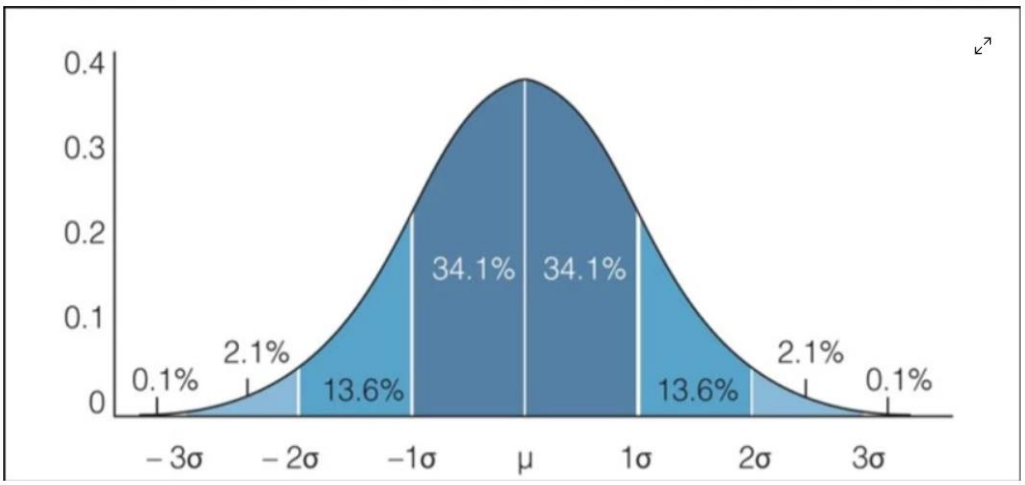
منحنى التوزيع الطبيعي ليه كذا خاصية :

- الوسط و الوسيط و ال mode
- المنحنى يتماثل نصف القيم على اليسار و نص القيم على اليمين
- المساحة تحت المنحنى هي 1

التوزيع الطبيعي بيتبع القاعدة التجريبية بحيث :

- 68% من البيانات تقع ضمن 1 الانحراف المعياري للمتوسط
- 95% من البيانات تقع ضمن انحرافين معياريين عن المتوسط
- 97% من البيانات تقع ضمن 3 انحرافات معيارية عن المتوسط

التوزيع الطبيعي هو التوزيع الاحتمالي الاكثر اهمية في الاحصاء لأن العديد من البيانات المستمرة في الطبيعة بتعرض المنحنى على شكل جرس لما بيتم رسمها بيانيا



Z statistic

لحساب احتمال وقوع حدث ما بنحتاج z statistic و صيغتها :

$$z = \frac{x - \mu}{\sigma}$$

اللي بتقوم بيه هنا هو توحيد المنحنى الطبيعي عن طريق تحريك المتوسط الى 0 و تحويل الانحراف المعياري الى 1

ال z statistic هو اساسا مسافة القيمة من المتوسط المحسوب بمصطلحات الانحراف المعياري

Central limit theorem

بتنص على ان عند التخطيط لتوزيع عينات من ال means ف المتوسط لمتوسط العينة هيكون مساوي لمتوسط السكان و توزيع العينات هيقترب من التوزيع الطبيعي

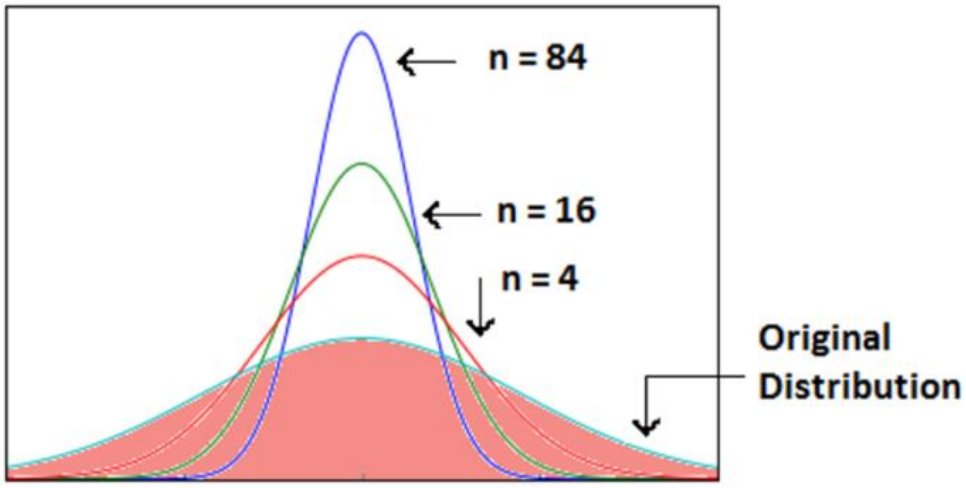
و هنا ممكن نلاحظ كذا حاجة :

- النظرية بتثبت الحدود المركزية بغض النظر عن شكل و نوع توزيع المجتمع سواء كان ثنائي الوسائط او منحرفا جهة اليمين

- يجب ان يكون عدد العينات كافيا لانتاج توزيع منحنى طبيعي بشكل مقنع و لازم يبقى في حذر للحفاظ على حجم العينة ثابتا لأن اي تغيير في حجم العينة سيغير شكل توزيع العينات

- كلما زاد حجم العينة انخفض الخطأ القياسي و زادت الدقة في تحديد متوسط المجتمع من متوسط العينة

مع زيادة حجم العينة يتقلص توزيع العينات من كلا الجانبين و دا هيمنحنا تقديرا افضل للإحصاء السكاني لأنه بيقع في مكان ما في منتصف توزيع العينات



Hypothesis testing

اختيار الفرضيات جزء من الاحصائيات اللي بتقوم فيها بوضع افتراضات تخص ال population parameter

اختيار الفرضيات بيشير الى اجراء مناسب من خلال تحليل عينة عشوائية من السكان لقبول او رفض الافتراض

اختيار الفرضيات هو طريقة لمحاولة فهم الافتراضات من خلال النظر في بيانات معينة

Type of Hypothesis

افضل طريقة لتحديد إذا كانت الفرضية الإحصائية صحيحة هي فحص المجتمع بأكمله و لأن دا غير عملى بنقوم بفحص عينة عشوائية من السكان

لو بيانات العينة غير متوافقة مع الفرضية الاحصائية فسيتم رفض الفرضية

في نوعين من الفرضيات الاحصائية :

- Null hypothesis

الفرضية بتقول ان عينة الملاحظات تنتج عن الصدفة فقط و بنرمز ليها ب Ho

- Alternative hypothesis

الفرضية بتقول ان ملاحظات العينة بتتأثر ببعض الاسباب غير العشوائية و بنرمز ليها ب Ha

Steps of Hypothesis Testing

عملية تحديد اذا كنت هترفض ال null hypothesis او لأ بناء على بيانات نموذجية العملية دي بنسميها اختبار الفرضيات و بتم من خلال اربع خطوات :

- state the hypothesis

بيتضمن ذكر الفرضيات الصفرية و البديلة و يجب ان يكون كلاهما متنافيا يعني لو احدهما صحيحا ف الاخر لازم يكون خاطئا

- formulate an analysis plan

يصف كيفية استخدام بيانات العينة لتقييم الفرضية الصفرية و التقييم دا في الغالب بيركز على احصائية اختيار الوحدة

- analyze sample data

بنبحث عن قيمة احصاء الاختبار و القيمة الاحتمالية الموضحة في خطة التحليل

- interpret results

بنطبق ال decision rule الموضحة في خطة التحليل لو قيمة احصاء الاختبار غير مرجحة بناء على الفرضية الصفرية بنقوم برفض الفرضية الصفرية

By ➔ Shorouk Eldeep