دا ملخص لكورس إحصاء من جوجل The Power of Statistics

محتويات الكورس:

1. الإحصاء الوصفيةDescriptive Statistics
2. الإحصاء الإستنتاجية Inferential Statistics
3. الإحتمالات Probability
4. أخذ العينات Sampling
5. مستويات اليقينConfidence Intervals
6. أختبار الفرضيات Hypothesis Testing

هشرح المفهوم بشكل بسيط وسلس مع استخدام صور وبعين هنكتب كود أزاي ننفذ دا في بايثون

الأحصاء نوعين: وصفيه ودي بتوصفلك الداتا وتديلك بيانا منها، او استنتاجية ودي بتقدر من خلال عينه مثلا تستنج معلومات عن الداتا كلها أو تتوقع مبيعات منتج

أولا الإحصاء الوصفيةDescriptive Statistics

دي وظيفتها تديلك المعلومات الرئيسية الأولية عن الداتا بتعتك ممكن في صورة أرقام أو رسوم بيانية

* زي المتوسط والوسيط القيم المتكرره وكل الحاجات الأوليه دي

Mean, Median, Mode, standard Division

بشكل منظم بقى هنقول إن النوع دا من الإحصاء بيدلك معلومات عن: مكان الداتا وتوزعي الداتا والقيم المركزية في الداتا

الـ Centeral Tendancy ودي فيها:

* Mean
* Median
* Mode

قواعد عامة لعلاقة الـ Mean and Median:

* لو الأتنين قربين من بعض يبقى توزيع الداتا طبيعي
* لو ال mean أكبر يبقى الداتا منحرفه شمال والعكس صحيح
* لو في فرق كبير قوي بينهم بقى أعرف ان عندك outliers

أنت كدا من رقمين قدرت تاخد فكرة عن توزيع الداتا وال ouliers

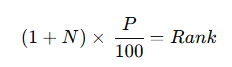
الـ Dispersion ودا فيه 3 قيم:

* Range ودا الفرق بين أكبر قيمة وقل قيمة في الداتا
* Standard Division ودا رقم بيقولنا القيم بتاعتنا قريبه من بعض ولا شبه بعض. بياخد المتوسط ويوقلك باقي القيم بتبعد عنه قد ايه، بيطلعلك رقم وبنقول ان القيم دي بتبعد وحدة معيارية والي بعد اتنين وحدة معياريه وهكذا
* Variance هو التباين ودا مربع الانحراف

ال Position ودا فيه 4 قيم:

1. Percentile

الولد دا percentile 90% يعني الولد دا درجته أكبر من 90% من زمايله. بنرتب القيم من الصغير للكبير ونستخدم القانون دا حيث ان الP هي النسبة المئوية الي انت عايز تشوف الي عندها جايب كام



1. Quantile

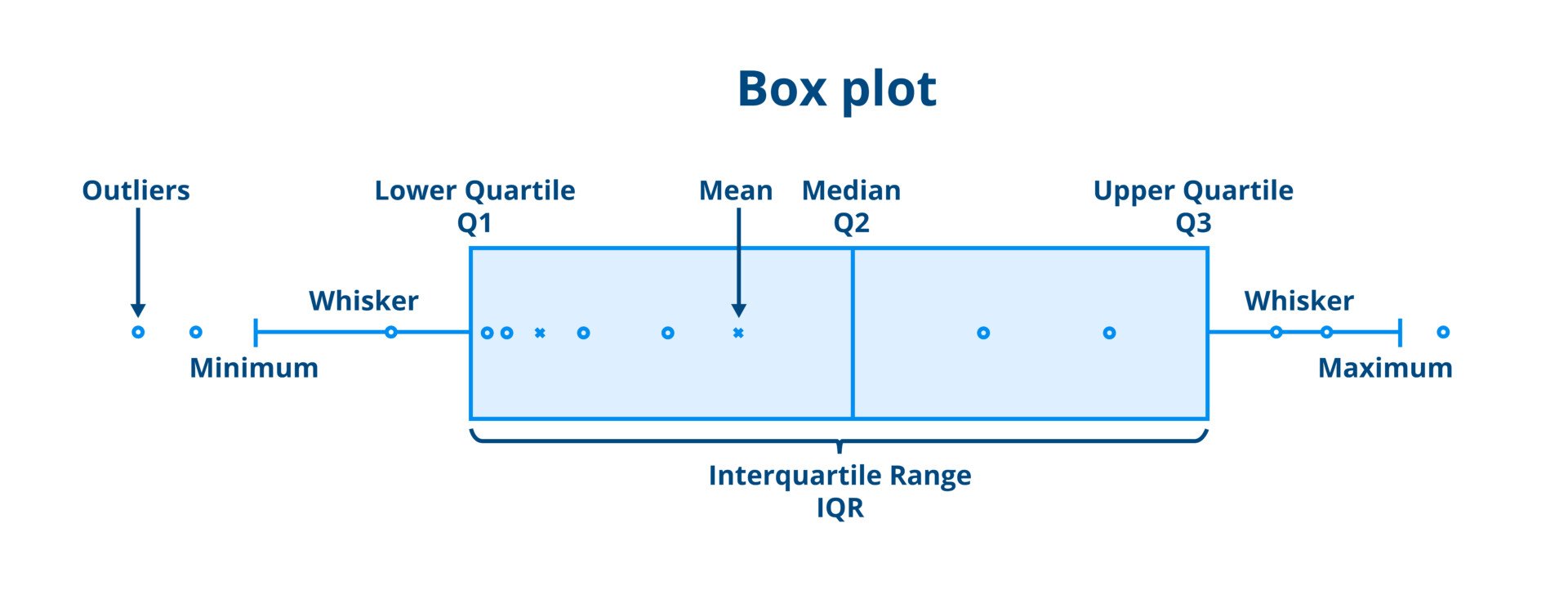
بقسم الداتا بتعتي على 4 أجزاء متساوية ودا بيقولي نسبة كام من الداتا تقع بين الرقمين دول

1. IQR interquartile range

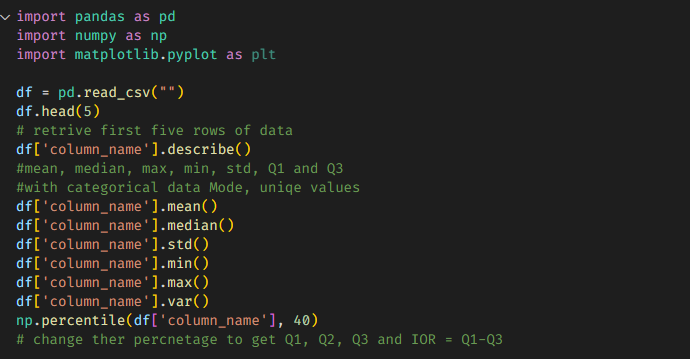
دا بيجبلك الفرف بين Q1 and Q3

1. Five number summary ودول هما
   * Minimum
   * Q1
   * median, Q2
   * Q3
   * maximum

والصورة دي فيها ملخص الكلام الي فوق



الخطوات على بايثون:



ثانيا الإحصاء الإستنتاجية Inferential Statistics

ودي فيها sampling, Confidence interval and Hypothesis test

بس قبل منشرح حاجة من دول لازم نعرف انها استنتاجية يعني في احتمال تطلع صح او غلط فلازم نعرف الأول عن الإحتمالات الي هي Probability

ثالثا ال Probability

بدرسها عشان اقيم انا متأكد قد ايه من الي بقوله ممكن يحصل بنسبة كام؟

هنقيمه ازاي؟ أقولك في نوعين من ال probability

* Objective بنقيم بالتجارب العملية بالأرقام بالإحصاء
* Subjective بنقيم بمزاجنا فوكك منها

ناخد ال objective probability ودي برضو يها نوعين:

* Classical بنستخدمها لما يكون كل الاحتمالات واردة بنفس النسبة equally likely وتعتمد على قوانين رياضية
* Empirical والكلمة دي معانها تجريبي يعني مبني على تجربة يعني نستخدمها لما الإحتمالات تكون غير معروفه او غير متساوية ودي بتعت بقى تجارب المعامل والقصص دي بيانات حقيقة يعني

لازم تعرف ان قيمة الاحتمالية بين ال 0 وال1 وممكن تضرب في 100 علشان تحولها نسبة %

لازم تعرف ان التجربة الي بنعملها ممكن يكون فيها حدث واحد او عدة احداث

لو حدث واحد نحسب احتماليته نقسم الحدث على عدد الاحداث الوارد حدوثه

مثال احسب احتمال ظهور ملك لو رميت عملة معدينة؟

الحدث واحد بس، عدد الاحداث كلها اتنين يبقى احتماليته؟ 50%

طيب لو عندي عدد من الأحداث؟ يبقى لازم لازم اعرف كام قاعدة كدا مرتبطة بالاحتمالية:

* Complement rule
* Addition rule
* Multiplication rule

وقبل ما تعرفهم لازم تعرف ان الاحداث نوعين:

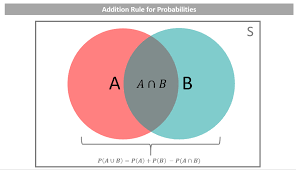
* Mutually Exclusive يعني مستحيل يحصلوا مع بعض، مستحيل يظهر ملك وكتابة في نفس الوقت
* Independent Events حدثين مفيش اعي علاقة بينهم رميت نرد ورميت عمله

لو عايزين نمثل الاحتمالية والحدث بشكل رياضي بنكتب كدا P(A) يعني احتمالية الحدث A.

وممكن أقول احتمالية انه ميكون A بكتبها كدا P(A’)

وممكن أكون أحتمالية الحدث A مع بمعلومية B كدا P(A|B)

الComplement rule بتقولك P(A) = 1-P(A’)

ال addition rule اختصارها في صورة P(A or B)=P(A)+P(B)−P(A and B)

مثال احتمال ظهور اتنين او اربعه لو رميت نرد؟ P(2 or 4) = 1/6 + 1/6

القاعدة التالتة multiplication rule

دي نسخدمها في نوعين من الاحداث ال independent يعني شربت قهوة احتمال الدنيا تمطر كام؟

P(A and B)=P(A)×P(B)

مثال ما احتمال ظهور ملك وظهور رقم 6 عند رمي نرد وعمله مع بعض؟

P(face and 6) =1/2 \* 1/6

والاحداث ال dependent مثال سحبت كارتين من كوتشينه بدون مترجع الكارت الاول، احتمال تطلع رقم 3 كام في المرتين؟

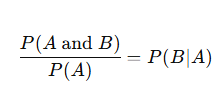
الكتشينه 52 فيها 4 ورقات برقم 3 يبقى احتمال رقم 3 اول مرة 4/52 وفي تاني مره 3/51

P(A and B) = 4/52 \* 3/51

ومن هذا المكان الجميل أحب اخدك على حاجة مهمه جدا اسمها Conditional Probability

ودي لما أقولك احتمال حدوث كذا بشرط انك تكون عارف ان الحاجة الفلانية حصلت P(B|A)

القانون بتاعها



مثال سريع

معايا كوتشينه ما هو احتمال اني اسحب ملك مع العلم اني الكروت الي معايا هي الحمراء بس

يبقى الحدث المعطي اني سحب لكروت الحمراء من الكوتشينه الي هما 26 دا يقولي ان فيهم كارتين ملك بس يبقى الاجابة

p = 2/26

مثال تاني فصل فيها 100 طالب 60 ولد 40 بنت 30 ولد بس بيلعلبوا كورة اخترت طالب ما احتمال انه يكون بيلعب كورة بشرط يكون ولد

الشرط الاخير هيخليلي العدد 60 وتوفر الشرطين مع بعض بيكونوا في 30 طالب يبقى النتيجة

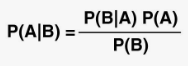
p = 30/60

الناس تسكت؟ والله قامت الدنا ولم تقعد تاني من ساعة مطلعت الاحتمال الشرطي دا

* قالك ممكن من القانون احسب انا احتمال الاتنين مع بعض



* وظهر العالم Bayes theorem بالقانون الجميل

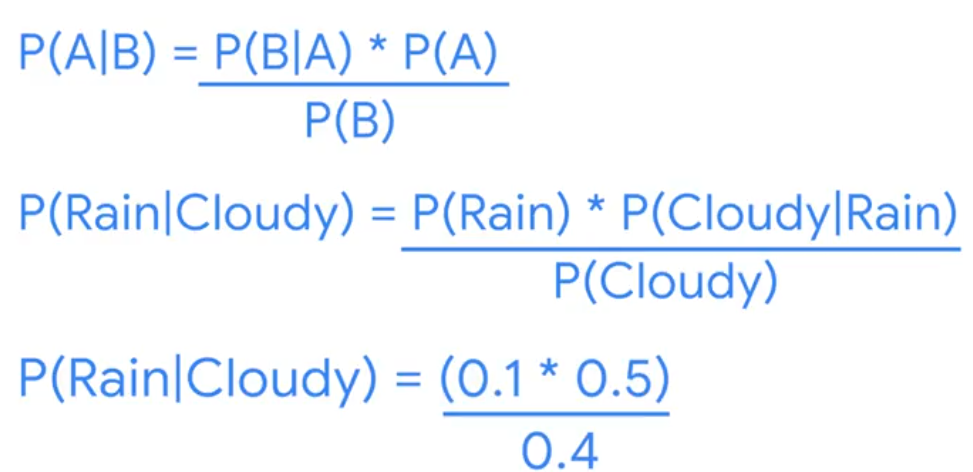


الراجل مكنش فاضي هو بيقولك ازاي لو حصل تغير في الحاجة الي انا بعتمد عليها ظهر دليل أو معلومة جديدة نعمل تحديث لقيمة الاحتمالية فبقى عندنا

* prior probabilityقبل منجمع الداتا P(A)
* posterior probability الجديدة P(A|B)
* والاحتمال بشكل عام P(B)
* واحتمال الدليل B لو حصل مع A اسمه P(B|A)

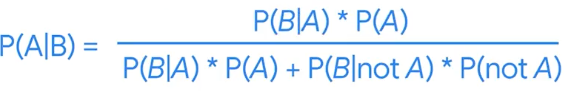
مثال عندك ميتنج وعايز تشوف الجو هيمطر ولا لا اصل الجو مغيم النهاردة زيادة

* في الوقت دا من السنة نسبة المطر 10% يبقى دي prior probability يعني P(A)
* احتمالية المطر في يوم غايم بتزيد 40% P(B)
* و 50% من الايام المطر بتبدا بيوم غايم P(B|A)



راح مسكتش عمل نسخة أطول كمان فكك ال P(B) قالك لو متعرفهاش

P(B) = P(B|A) \* P(A) + P(B | A’) \* P(A’)



مثال تقييم اختبار الكشف عن مرض

* A معاناها المريض فعلا مصاب
* B الاخبار طلع ايجابي

بيقولك 1% من السكان عندهم المرض P(A)

احتمال ان الاختبار يطلع ايجابي وانت عند المرض 95% P(B|A)

احتمال يطلع ايجابي وانت معندكش المرض 2% P(B| A’)

السؤال ما احتمال ان المريض يكون مصاب والاختبار ايجابي فعلا P(A|B) خلي بالك احنا مش عندنا B الي هي احتمالية الاختبار يطلع ايجابي فهنجيبها بالصيغة المطولة

P(A|B) = (.95 \* .1)/((.95 \* .1)+(.2\*.99))

موضوع أخير مهم جداً هو الـ Probability Distribution

هنا انا بحاول اعرف احتمالية حدوث كل الأحدث وبعبر عن كل حدث بحاجة اسمها Random Variable مثال رميت نرد فيه 6 احتمالات عايز 6 random var زي[1-2-3-4-5-6] وبحط قصاد كل واحد احتماليته

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 6 | 5 | 4 | 3 | 2 | 1 | rv |
| 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | Probability |

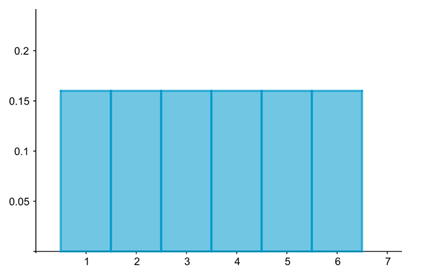
الRandom Variable في نوعين منها Discrete and Continuous

الdiscrete ارقام منفصله قائمة بذاتها مفيهاش ارقام لا نهائية زي مثال النرد

ال continuous ارقام لا نهائية مثال حساب متوسط طول البشر هتلاقي 2.80 و2.90 وبينهم ارقام لا نهائية من البشر

لو جينا بقى نوزع الداتا دي على شكل بياني هتطلع ازاي؟

مثال الجدول الي عملناه فوق عن النرد:



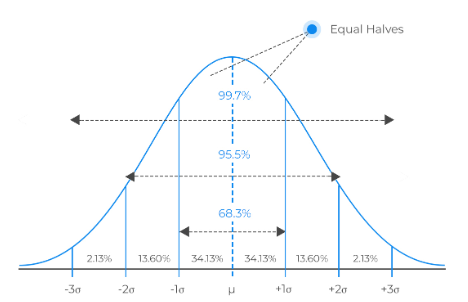
كل نوع من الـRandom Variable ليه اشكال توزيع معينه هعمل جدول يلخص

ال Discrete

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **التوزيع** |  | **شروط التجربة** |  |  |
| **Binomial** | **التوزيع الثنائي يعني النتيجة حاجة من اتنين، يعني mutualy exclusive** | **تكرر عدة مرات، يكون ليها احتمالين فقط ليهم نفس الاحتمالية** | **في قانوان تطلع بيه الاحتمالية متشغلش بالك بيه** | **رمص عملة 5 مرات احتمال تطلع ملك ؟** |
| **Bernoil** | **زي باموميال بس التجربة بتحصل مرة واحدة بس** |  |  |  |
| **Poisson** | **بستخدمه لو عايز اعرف احتمال حدوث شيء معين في مدة زمنيه معينه** | **عدد الاحداث يمكن عده، المتوسط في المدة الزمنية معروف، كل حدث مستقل** | **ليها قانون معين** | **عدد المكالمات في الدقيقة، كم زائر في ساعة** |

ال Continuous

هنشرح ال Normal او Gaussian توزيع الداتا بيكون متساوي حوالين الMean على شكل جرس



في حاجة اسمها ال empirical rule متعلقة بالشكل دا

approximately 68% of data points fall within one standard deviation of the mean, about 95% within two standard deviations, and about 99.7% within three standard deviations

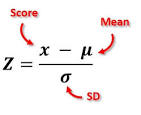
أحيانا بيكون عندي داتا مختلفه وعايز اشوف العلاقة بينهم او عايز اشوف العلاقة بين قيم الداتا الواحدة هنا بلجأ لحاجة اسمها Z-Score

لو معايا داتا بوينت دي بتقيس كم عدد الانحرافات المعياريه اقل او اعلى المتوسط بتاع الداتا كلها مش عينه.

الفكرة هنا اني صعب اقارن بين قيم مختلفه فبلجأ اني اوحد الداتا المختلفة لمتوسط صفر وانحراف 1 هي مشتقة من التوزيع الطبيعي

بطرح القيمة من المتوسط واقسمها على الانحراف

بتطلع قيمة الz وتطلع الاحتمالية بتعتها من جدول التوزيع الطبيعي z-table



Z-Score أداة قوية لفهم مكان القيمة بالنسبة للتوزيع، واكتشاف القيم الشاذة، وحساب الاحتمالات، ومقارنة نتائج من توزيعات مختلفة، وتوحيد البيانات.

لو Z-Score كبير جدًا (مثلاً ±3)، نعتبر القيمة دي شاذة أو Outlier.

متوسط أطوال الرجال = 175 سم، الانحراف المعياري = 8 سم. رجل طوله 190 سم. Z=1.87

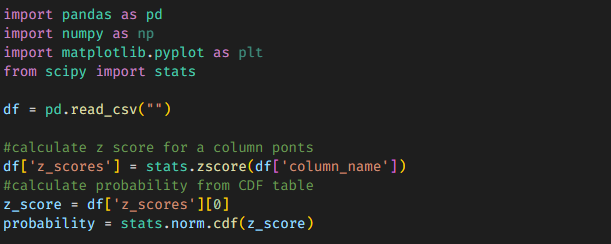
ده معناه إنه أطول من المتوسط بحوالي 1.9 انحراف معياري بناخد الرقم دا ونروح للجدول يطلع 0.9699 يعني بنسبة 97% تقريبا

لو عندك امتحانين الأول متسوط الدرجات 70 وانحراف 10 وانت جبت 85 والتاني 200 انحراف 20 ودرجتك 240 احسب ال z الأول يطلع 1.5 والتاني 1.6 عارف لو القيمة بالسالب يبقى انت كدا اقل من زملائك حتى بس في الاختبار الاول انت منحرف عنهم ب1.6

حساب احتمال حدوث شيء معين احتمال طالب يجيب 75 لو المتوسط 80 والانحراف 10 الz هتكون -0.5 بنروج الجدول CDF تلاقي النتيجة 0.3085 يعني 30%

كمان نسخدمها علشان نعمل standardization يعني كل عمود يبقى متوسطه 0 وانحرافه 1، علشان النماذج تتعامل مع البيانات بشكل عادل.

كمان اختبارات الفرضيات (Hypothesis Testing)



رابعا الـSampling

انواع العينات الي بناخدها فين نوعين عامين Prbability sampling methods and non probability

ال Probability Methods دي 4

|  |  |
| --- | --- |
| simple random اختيار عشوائي وكل الاحداث ليها نفس الفرصة |  |
| stratified random بقسمهم مجموعات حسب السن او الجنس واختار عشوائي من كل مجموعة |  |
| cluster random قسمهم مجموعات وخ عدة مجموعات كاملة |  |
| systemic number رتبهم في تسلسل وقول هاخد الرقم الخامس دايما وهبدأ من رقم كذا |  |

التوزيع العشوائي بيضمنلي ان العينة مفيهاش اي انحياز Bias

النوع التاني non probability methods رخيص وسريع بس مش مضمون تستخدمها لما ميكون شعندك وقت تنفعك لو عايز تستشكشف بشكل مبدئي مش تعمل استنتاج نهائي.

1. convenience بشوف مين الي اقدر اوصله واتعامل معاه واختاره
2. voluntary response بشوف مين عايز يتطوع
3. snowball تسويق شبكلي كدابيختار ناس تقول لناس تعالوا
4. purposive بختارهم بالغر الي انا عايزه قايمة كلاب اختار الي جايبين درجات اعلى

نبدأ نشتغل على العينة بقى احصائيا

هنا بروح اخد قيمة احصائية sample statistic وابدا اششتغل عليها زي ال mean مثلا، اهو السطر دا اسمه Point estimate

واخده واروح اعمله sample distribution ودا معناه اني بشوف احتمالية التوزيع بتاع الsample statistic

عايزك تعرف انك بتاخد اكتر من سامبل علشان تعمل الدستربيوشن دا

كل مالعينة تكبر كل ما القيمة الي بقيسها تكون قريبة من قيمة الداتا الكليه population

في حاجة اسمها standard error هو الفرق بين Population

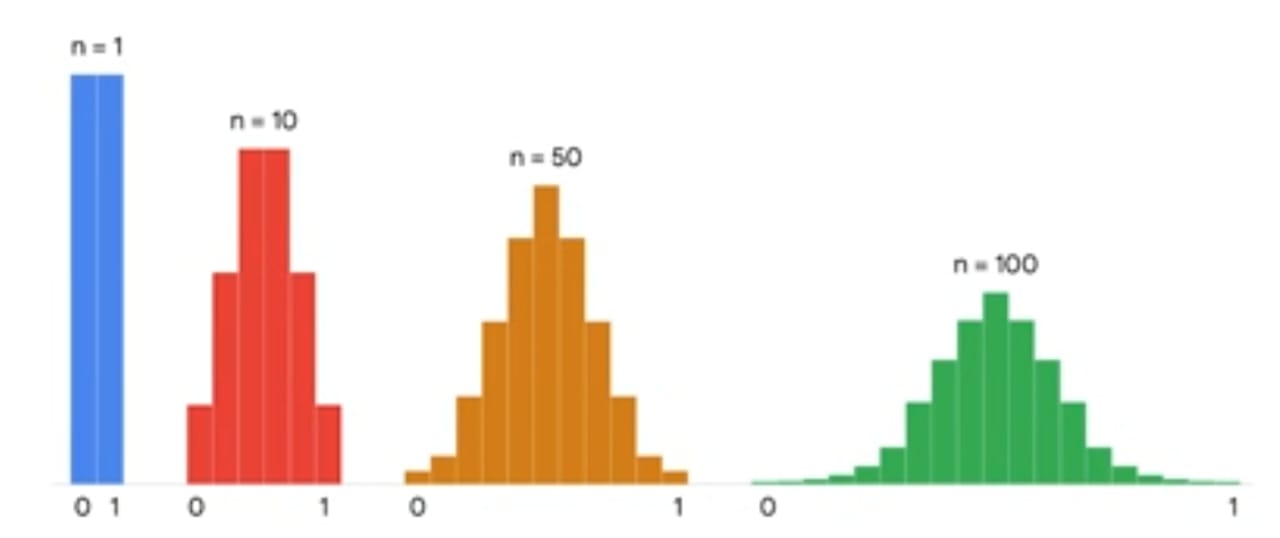
لو عايز احسب الانحراف المعياري standard error = انحراف العينة/جذر عدد العينة

**SD** مدى تباعد القيم عن المتوسط.

**SE**  مدى دقة المتوسط كتمثيل للمجتمع.

أخر حاجة في الباب دا نظرية مهمه اسمها Central Limit Theorem

بتقولك التوزيع الاحتمالي بتاع متوسط العينة بيقرب انه يكون توزيع طبيعي bell shape كل ما تزود حجم العينة



**الأساس للعديد من الاختبارات الإحصائية** (t-test, z-test…).، **استخدام التوزيع الطبيعي حتى مع بيانات غير طبيعية** إذا أخذنا عينات كبيرة. تساعدنا على **تقدير المجتمع** من خلال العينات.

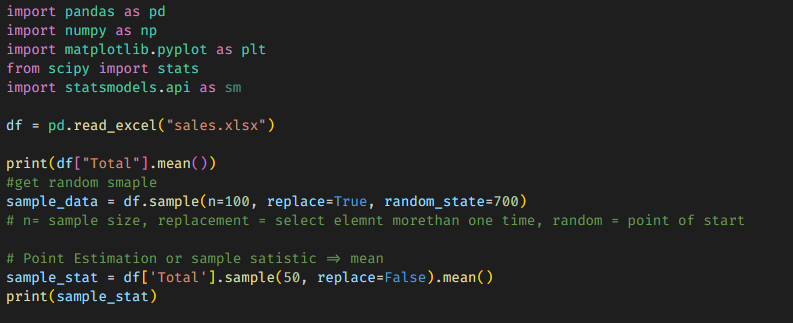
نقطة اخيرة قيمة احصائية اسمها Proportion نسبة الناس ضمن ناس كتير فشخ بس دول ليها خصائص معينه

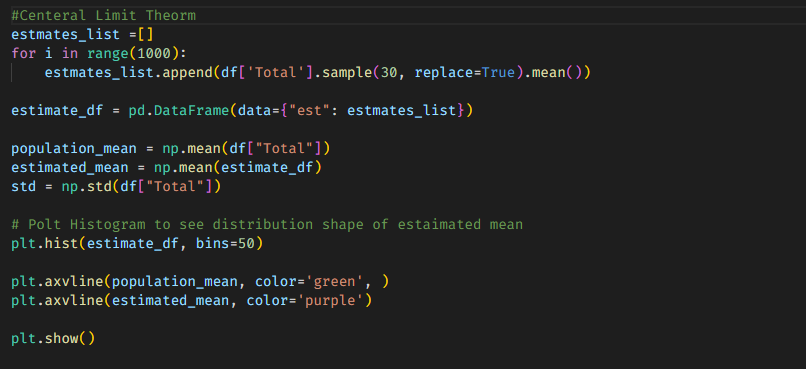
وفي قانون خاص بكل قيمة احصائية تقدر تحسب بيه Standard Error بتاع ال Proportion ودايما فرق بين

ال Parameter قيمة احصائية تعبر عن ال population

وال Statistic قيمة احصائة تعبر عن ال sample

والقوانين بتختلف من مكان لأخر





رابعاُ فترة الثقة Confidence Intervals

هو رنج معين من القيم نعتقد ان القيمة الي بندرو عليها موجودة خلاله بنسبة ثقة معينة

عكس الpoint estimation هو قيمة واحدة مش رينج

نسبة ثقة معينة الي ف التعريف اسمها confidence level ودي بتكون 95% دا الي مشهور بس ممكن نغيرها حسب احتياجنا

ال interval هو sample statistic -/+ margin of error

ايه هو margin error هو اقصى فرق متوقع بين pop parameter and sample estimate

Margin Error = z-score \* Stander Error

الz score بيتحسب هنا من خلال الConf Leve وكل مستوى ليه قيمة z ثابته

|  |  |  |  |
| --- | --- | --- | --- |
| Cl | 90 | 95 | 99 |
| Z score | 1.645 | 1.96 | 2.58 |

يبقى الخطوات كالتالي:

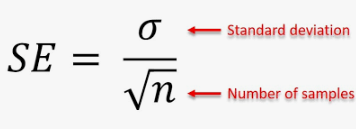
1. Choose sample statistic
2. choose confidence level
3. calculate margin of error
4. calculate interval upper and lower

مثال: معايا 10000 بريق عايز اعرف متسوط اوزانهم ايه؟

هاخد عينه 50 بطريق وأطلع المتوسط بتاع الوزن sample statistic = 30والانحراف 3

عند مستوى ثقة 95%

هامش الخطأ = زد سكور \* Stand Error



ف ال se = 1.3/جذر50 = 0.42

يبقى هامش الخطأ = 0.42 \* 1.96 = 0.8232

يبقى الحد الأعلى = المتوسط + هامش الخطا والأدني ناقص هامش الخطاء

فبقول بنسبة 95% متوسط اوزان البطاريق ال10000 هتطلع بين 30.823 و 29.28

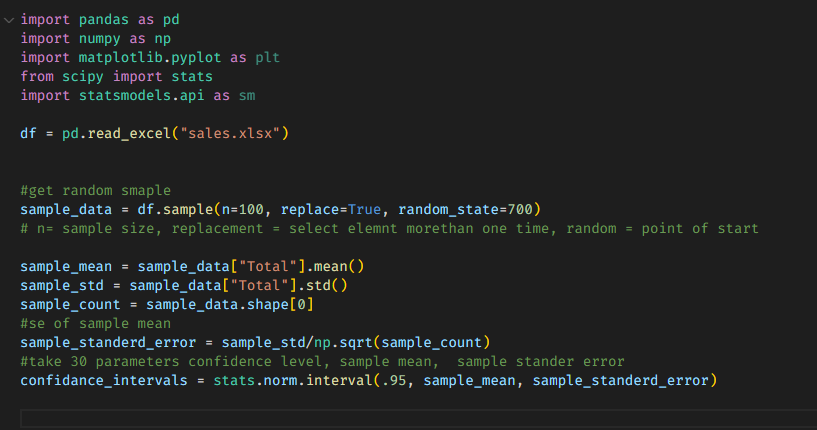
ملحوظة مهمه جداً

في الإحصاء بيحبو يتعاملوا مع رقم يعبر عن باقي العينة او المجموعه الكبيرة، ولذلك لجأو انهم يخترعوا حاجات زي z-score وعلشان الدقة تبقى أكتر اخترعوا حاجات تانية مناسبة لحالات تانية الحاجات دي اسمها statitical test امثله عليها z-test, t-test, chi square, f test

الz مثلا بيقارن متوسطات عينات كبيرة

الt نفس الفكرة بس لو العينة صغير >= 30

وهكذا يبقى دا رقم بجيبه يعبرلي عن حاجة معينه علشان اقدر اخد الرقم دا اقارنه بحاجات تانية فيه تفاصيل كبيرة وقوانين اعرفها عند الحاجة وخلاص



الفصل الأخير أختبار الفرضيات Hypothesis test

قصته اني بشوف هل بيانات العينة بتديني دليل كافي اني اقبل او ارفض فرضية معينه عن الpopulation ؟

وبنقيم الدليل بكلمة Statically significant ومعناه ان نتيجة الاختبار او التجربة مش جايه بالصدفه لا في دليل مقنع فعلا عليها

فانا عندي فرضيتين

null hypothesis Ho ودا الي بيمثل الوضع الحالي يعني عدم وجود فرق ودي ديما صح طالما مفيش دليل على عكسها

alternative hypothesis Ha ودي الفرضية الجديدة بتقول في فرق ودي غلط طالما مفيش دليل

وممكن تكوين لا تساوي او اكبر من او اصغر من

وارد نعمل خطأ في الحكم على نتيجة اختبار الفرضية يعني مثلا

وارد نرفض الnull وهي اصلا صح فدا اسمها False – Positive او Type I Error

وارد نفشل في رفضها نقبلها يعني وهي اصلا غلط False – Negative او Type II Error

واحنا بشر فبنحط لنفسها هامش خطا وبنسميه Significance level

ودا معناه احتمالية رفض الnull مع انها true وغالبا بيكون 5% وبيتغير طبعاً حسب الحاجة

أوك يا سيدي الفاضل ندخل في الارقام والاحصاء تحب الارقام، فاكر لما قولنا الاختبارات الاحصائية معمولة علشان تعبرلي عن بيانات العينة برقم واحد؟ فاحنا هنستخدمها هنا برضو

طلعت ال z score مثلا ها اعرف ازاي بقى ان الفرضة بتعتي صدفة ولا عليها دليل ؟؟

هنا يظهرلنا حاجة جديدة اسمها p-value

لو الـ p-value صغير جدًا (مثلاً أقل من ٥%): نقول «الظاهر في حاجة مش طبيعية» ونرفض الفرض الأصلي. لو الـ p-value مش صغير: نقول «النتيجة ممكن تكون بالصدفة» ونسيب الفرض الأصلي.

يعني بيقيس مدى توافق البيانات مع H0

بنقارنها مع الSignificance level ولو هي اصغر منه يبقى نرفض ال Ho ولو أكبر منه يبقى بنقول failed to reject null

حاجة كمان في نوعين من الاختبار

one sample ودا لما اكون عايز اعرف هو البراميتر زي المتوسط او النسبة يساوي قيمة معينة ولا لا

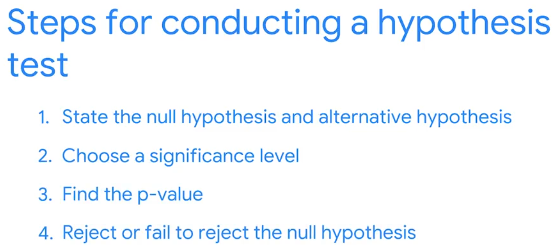
زي اني عايز اشوف متوسط ايرادات الشركة يساوي التارج بتاعنا ولا لا

مثال شركة توصيل بتقول ان متوسط وقت توصيل الاوردرات 40 د والانحراف 5

بعدين الشركة عملت برنامج تدريب وتطوير علشان تقلل الوقت دا

راح خدوا عينة من الي خدوا تدريب وليكن خمسين ولقيو المتوسط 38 والانحراف 5

عايز بقى نشوف الفرق في الوقت دا statically significant ولا هو جاي صدفه

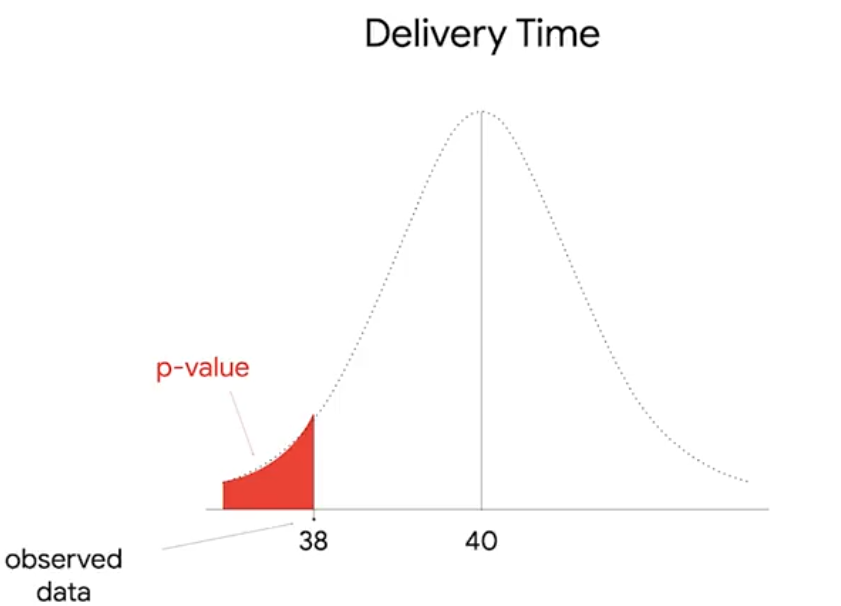
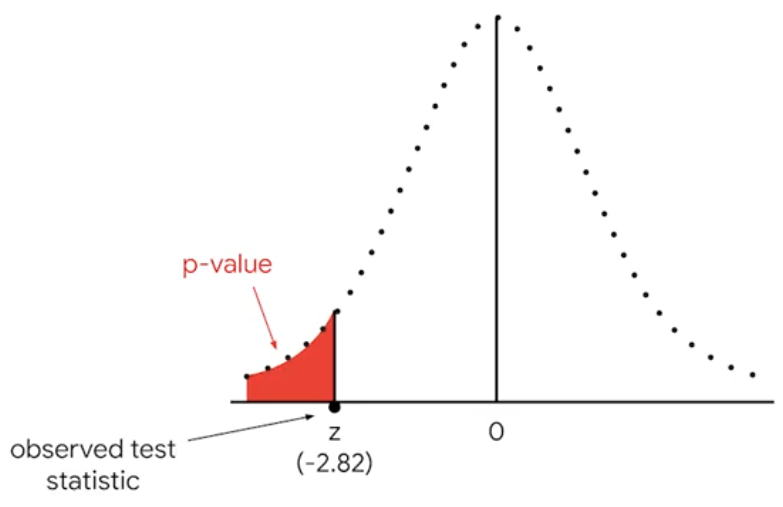


H0 التوصيل 40 دقيقة متوسط

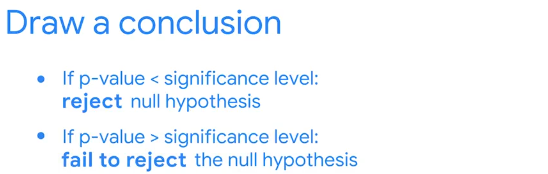
Ha التوصيل اقل من 40 د

ألsig level هنقول انه 5%

الp value بتتحسب الكمبيوتر لكن هي بتتحسب من test statistic

 بعدين بتحسب ال z طلعت -2.82

ولما حسبنا الp بالكمبوتر طلعت 0.23% فهي اقل من 5% فهنرفض ال H0



ألنوع التاني Two sample test بقارن قيمتين هل هما زي بعض ولا لا

مثال شركة هتغير تصميم صفحة معينة وعايز تشوف تاثير دا على المدة الي العماء بيقعدوها على الموقع خلي بالك انا معرفش الوقت المتوسط بتاع الناس بيقعدوا ع الموقع كام وهنا هلجا لt test علشان مبتطلبش القيمة دي

العينة الاولي حجم 40 متوسط 300 ث انحراف 18.5 ث

العينة التانية حجم 38 متوسط 305 ث انحراف 16.7 ث

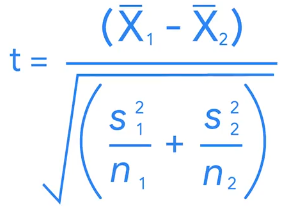
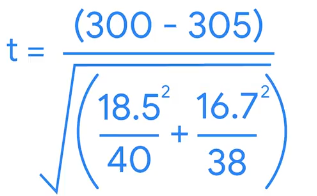
هنعمل اختبار

H0 بتوقل انه مفيش تاثير بين التصميم القديم والجديد

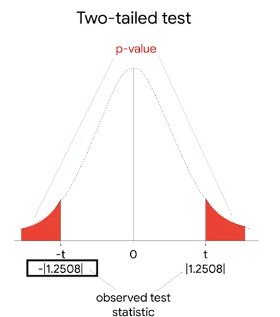
Ha بتول انه في فرق في متوسط الوقت والتصميم الجديد ليه تاثير

الsig le 5%

احسب ال p value على الكمبيوتر بس لازم الاول تحسب الt test



تطلع قيمة ال t تساوي -1.2508



احسب ال p هتطلع 21.48% هي بتساوي مساحة المنطقتين الحمر

الp اكبر من ال significance level دامعناه اني فشلت ارفض ال Ho ويبدوا ان الفرق دا جاي بالصدفه

