Foundations of Data Science: Randomness, sampling and simulation – Introduction to statistical inference

Where are we in the course?

I. About data: collection, representation, wrangling, exploration, visualisation and descriptive stats

II. Intro to Machine Learning

III. Linear models

IV. Statistical Inference

V. Regression and inference

Descriptive statistics

STATISTICAL DYO interity Tecount of The Statistical account of Scotland, commenced in May 1190, and was completed in 1888.

A the statistical account of Scotland commerced in Stop 1190, and was completed in 1840. The Publication of the corrected Oranty Reports, commenced in Smel1905 and was completed in 1814. 3 The General Report of Scotl<u>and commenced in 1811, and was comp</u>leted in 1814.....

To complete these several undertakings required in all a period of about Twenty four Years, and the assistance of above one Thousand todividuals.

LAUSDEO FINITUM.

down their land m good condition for grafs, all I sh Lime, is, that by it we can produce good Grops of rough A, these last will not grow in this country, the crop a but often productive of of well got, the strow is excelled The sheep . would the farmers consult their own Intere Their lands with grafo the second or third crop, the Hay more than compensate them, besides leaving the Land in crops, but men seldom forgoe a present profit for for 12 - Stough: gates in the bounty of M Lothian by wh is determined at 15 35-2 ------226.11.3 Do Setherh shire ----- 20 Horses Mid Lothian - - - - 182 Do Selhurt shere ----Black Gattle Med Lothean +2.90 Do Delhurh Ahre 199 Carto Med Lothean -----De Selhurt shore ------ 30 considered is Sheep, for which nature seems chiefly this part of the country, If we examine the theep . The parish, they seem orriginaly to have been of th

https://stataccscot.edina.ac.uk

Inferential statistics

The process of drawing conclusions about quanties that are not observed.

E.g. Wildcats



Wikipedia, Peter Trimming, CC BY 2.0

We observe the mean of a sample

We infer the mean of the population

E.g. "Manuscript on Deciphering Cryptographic Messages" Al-Kindi, 9th Century, Baghdad

دا. سعالده على دالدمر وصف ما كلومالغن العرد مرد الماري مراجع مراسع له مر ما عام الروسان سعد مع وصنع من كلوم المن مع العرف ما عوال هر وسعاد وللمعل - مار مسلما الإلطرو بلحويل والماري الوضية مع الما مسيح وما در وسعاد و مار مع ما مار مسلما المدول والماري المع مع المار المساعد والروسل المحرور المعلى - مراكحها المرادة والماء المريز وكل والمد والرك والمع السر المراد المساحة الراس - مسم المراد و والما المريز وكل والمد والرك والمع وحد والمراس الماء المساحة الراس السم والمراد و والما المريز وعسار الطرو ما المد والمراح المعر والمحد و

فرااداد والجعدانه ودائعال موصلوا يستعلم مرعد والجد ع

سمايد الم مسموالي مرالي معدم وسالد الاست مرالي مسمور الدي استوام العرج المواسر فعن معالد فعل وفر علما مرزيس وكال موديد الحسل المستالم مارس التركيم ماد ولت ورالا ما مرزيس وكال موديد الدرسيل سدال المناص العوا عنه الله اسلود فراه ما ورود من الغل المنه مراليو ورسيور المعالم مراسم عنه الله اسلود فرالا ما وجوهما و ولعروالما العقر والمعلي المراجع المراجع الرامعاد وسعول واراله ما وجوهما و ولعروالما العقر والما العوالي م الرامعاد وسعول واراله ما وجوهما و ولعروالما العقر والمعلي المراجع المعالي م الرامعاد وسعول واراله ما وجوهما و ولعروالما العقر والمعلي المراجع العروم

We infer the meaning of the messages

Inferential statistics tasks

- 1. Estimation
- 2. Hypothesis testing
- 3. Comparing two samples (A/B testing)

Inferential statistics tasks: Estimation $[304g], 336g] \in CI$. $\hat{\mu} = \bar{\chi} = 320g \pm 16g N = 20$







Peter Trimming, CC BY 2.0, Wikipedia

Point estimates

Confidence intervals: how confident are we in the estimate?

Inferential statistics tasks: Hypothesis testing

Yes/no questions: E.g. 1: "Is Chocolate good for you"

E.g. 2: Swain versus Alabama (1965). Is this jury selection procedure biased?

Population of Jury panel of selection Alabama : 001 2690 Black <u>8</u> Black and 92 Non-black 74% Nonblack

Inferential statistics tasks: Comparing two samples (A/B testing) E.q. 1. Is a vaccine better than a placebo? E.g. 2. Does providing recommendations lead to more sales? Treatment (Recommand. ations) 989 Control (no recommendation)

Two approaches to statistical inference

- 1. Computational: "Statistical simulations"
 - + Few assumptions => can be applied to many situations
 - + Little theory required
 - + Hopefully intuitive
 - Can be compute-intensive
- 2. Mathematical: Statistical theory
 - + Not compute-intensive
 - + Standard in scientific literature
 - Can depend on assumptions that aren't true (e.g. normal distributions)

Plan for statistical inference

- 1. Randomness, sampling and simulations (S1 Week 10)
- 2. Estimation, including confidence intervals (S1 Week 11)
- 3. Hypothesis testing (S2 Week 1)
- 4. Logistic regression (S2 Week 1)
- 5. A/B testing (S2 Week 2)

How can we address these questions?

- 1. What is the mean and median age of the population of all 2p and 10p coins in circulation?
- 2. Are tosses of 2p and 10p coins biased, i.e. is the probability of heads or tails different from 1/2 ?





Old style

New style

Let's get sampling!

- 1. Go to the form at the right
- 2. Record the
 - denomination (2p/10p)
 - style (old/new)
 - year
- 3. Toss the coin 8 times and record the results
- 4. Submit the form

Coin tossing data



https://forms.office.com/e/SKNgiQmB4N

Results

How certain are we that the mean year is what we compute?

Do we think that the coins are biased or not?