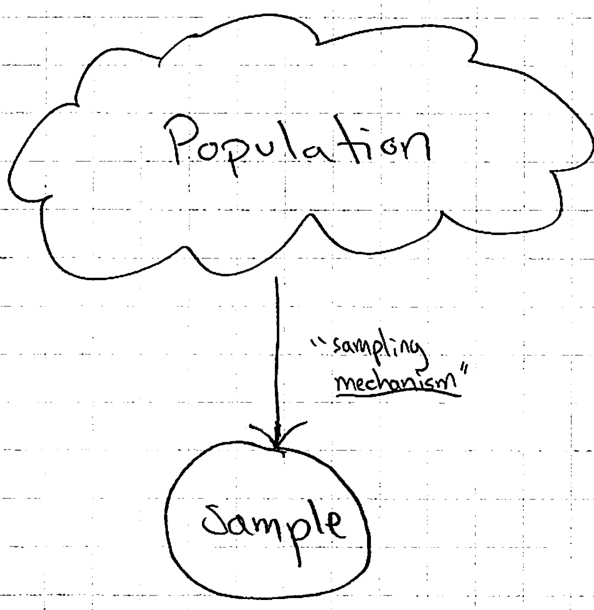
"Central Dogma" of Statistics



set of all data of interest
- heights of all humans
- ratings from all buyers
- often theoretical/abstract quantity due to its size

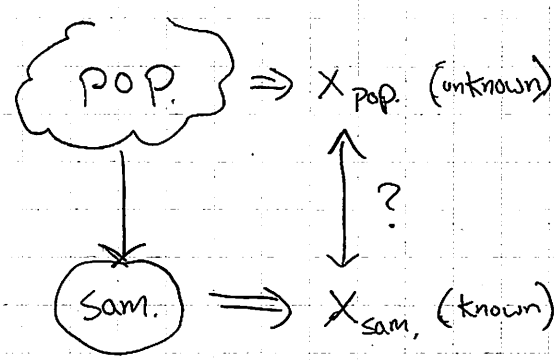
subset of data collected from pop.
- set of observations

Sampling mechanism ⇒ experimental action, how data collected
- survey/interview process
- field sample collection
-

Question: How representative is the sample (of the population)?
Is the mechanism biased in some way?

Two main tasks: Inference and Prediction

Inference: what can we conclude about pop given sample?

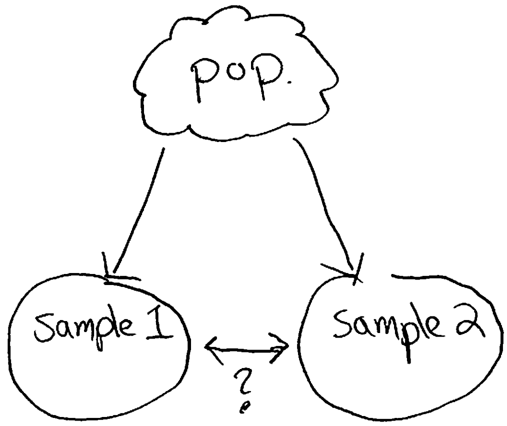


X: summary statistic

Ex: mean height of people in sample vs. population

⇒ To understand how $X_{sam.}$ relates to X_{pop} will require understanding the sampling mechanism, size of the sample, and properties of the statistic(s).

Prediction What can we conclude about a new sample from our first sample?



Machine Learning/Statistical Learning task:

$\Rightarrow$ Use sample 1 to build a predictive model f of summary statistic and apply to sample 2

$\Rightarrow$ "training" $\Leftarrow$

$X_{sam1} \Leftrightarrow f(\text{"sample 1"})$, learn f

$f(\text{"sample2"}) = ?$

Prediction is often purely computational, and does not consider the population or sample mechanism.

Ex: Predicting whether or not a photo contains a cat

pop - all photos ever taken

sample 1 - collection of known photos
$X_{sam1} = \{\text{"cat"}, \text{"not cat"}\}$ } training data used to learn f

sample 2 - 1+ new photos...

f - ML/SL method (conv. neural net?) that takes a photo as input and returns "cat" or "not cat" $\Rightarrow$ f is a classifier.

$\Rightarrow$ Inference and Prediction are closely related.
- often inference methods are also good at prediction
- understanding sample uncertainty helps w/ both Inf. and Pred.

Statistics and Probability statistics is not part of mathematics. It is a mathematical science that uses mathematics to quantify and understand data and data collection processes (i.e. experiments and observational studies)

modeling the mechanism as a random process $\Rightarrow$ probability is the primary mathematical tool.

Ex. Each member of the pop. is equally likely to be chosen by the mechanism to be put in the sample.

$\Rightarrow$ model the mechanism as an "coin flip"
↳ approximation, but maybe useful

We can quantify our uncertainty in how sam. relates to pop. by modeling that uncertainty w/ randomness. $\Rightarrow$ probability

Review Probability

Variables - property/descriptor that can take on multiple values

$\Rightarrow$ variable is a question, its value is the answer

Ex How old is this individual? 38 years old.
$\Rightarrow$ variable: age, value: 38

- The probability that variable X takes value x is $P(X=x)$ sometimes shortened to $P(x)$.
- Prob. of multiple values at once: $P(X=x, Y=y)$
- Variables can be discrete (categorical) $\Rightarrow$ take on one of a finite or countably infinite set of values in any range or continuous $\Rightarrow$ take on one of an infinite set of values on a continuous scale (for any two values, there is a third value between them).

Events assignment of value(s) to variable(s)
"$X=1$", "$X=1$ or $X=2$", "$X=1$ and $Y=3$", etc.
Any declarative (true/false) statement is an event.
$\Rightarrow$ Different from everyday definition. "Joe is 20 y.o." vs. "Joe turned 20 y.o."

Conditional Probability: Prob event A occurs, given that we know some other event B has occurred. $\Rightarrow$ cond. prob of A given B.

Prob. $X=x$ given $Y=y$ is written $P(X=x|Y=y)$ or sometimes just $P(x|y)$.

Ex. prob. you have the flu, vs. prob. you have the flu given your temp. is 102°F (38.9°C)

Independence - prob of one event does not change w/ obs. of another event.

- Ex. Prob. you have the flu given your friend Joe is 20 y.o.
- Events A and B are independent if $P(A|B) = P(A)$. If not they are dependent. Likewise $P(X=x|Y=y) = P(X=x)$ for variables.
- Indep. and dep. are symmetric. If A depends on B then B also depends on A; A indep of B, then B is also indep. of A.
- Dependence $\neq$ causality!

Probability distributions - Prob. distr. for a variable X is the set of probabilities assigned to each possible value of X.

Ex X can be 1, 2, 3. A prob. distr. for X would be $P(X=1) = \frac{1}{2}$, $P(X=2) = \frac{1}{4}$, $P(X=3) = \frac{1}{4}$.

Probabilities must be between 0 and 1 and sum to 1. Sometimes called a p.m.f. (prob. mass function)

If X is continuous we instead define a density function f. Prob X is between a and b is the area under the curve $f(x)$ between $x=a$ and $x=b$: $\int_a^b f(x)dx = Pr(a<x<b)$

$\rightarrow f(x)$ must also satisfy $\int_{-\infty}^{\infty} f(x)dx = 1$.

Joint distributions describe probability distributions for sets of variables.

Ex $P(X=1, Y=1) = \frac{1}{5}$, $P(X=1, Y=2) = \frac{1}{10}$, etc. must also sum to 1.

$\downarrow$

all combinations of values.

Probabilistic Truths

$P(A \text{ or } B) = P(A) + P(B)$ if A and B are <u>mutually exclusive</u>.*

otherwise, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

For <u>any</u> events A and B: $P(A) = P(A,B) + P(A, \text{"not } B\text{"})$
b/c "A and B" and "A and 'not B'" are mutually exclusive.

More generally, for any set of events $B_1, B_2, \dots, B_n$ such that exactly one of B_i must be true:

(* m.e. implies dependent!)

$$P(A) = P(A,B_1) + P(A,B_2) + \cdots + P(A,B_n)$$

This is the <u>law of total probability</u>.

- Calculating $P(A)$ by summing its prob.s over all B_i's is called <u>marginalizing</u> and $P(A)$ is called the marginal distribution of A.

If we know prob of B and prob of A given B, we can determine prob. of A and B

$$P(A,B) = P(A|B) \cdot P(B) \longrightarrow P(A|B) = \frac{P(A,B)}{P(B)}$$

(formal definition of cond. prob.) note if $P(B)=0$

⇒ this also gives us a numerical representation of independence: $P(A,B) = P(A) \cdot P(B) \Rightarrow P(A|B) = P(A) = \frac{P(A,B)}{P(B)} \rightarrow P(A,B) = P(A) \cdot P(B)$

⇒ And combining this w/ symmetry relation $P(A,B) = P(B,A)$ gives us something very important:

$$P(A,B) = \underbrace{P(A|B) \cdot P(B)} = P(B,A) = \underbrace{P(B|A)\,P(A)}$$

$$P(A|B)P(B) = P(B|A)P(A)$$

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

Bayes' Rule (Bayes' Thm)

[We will use Bayes' Thm. in the future, quite a bit]

weighted sum of cond. prob.

Lastly, $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$
combining law of total prob. w/ conditional prob.s

<u>Statistics</u> A statistic is a (numerical) measure of a "feature" of a probability distribution

- Expected Value aka "mean" (expectation) can be used for variables w/ numerical values.

The expected value $E[x]$ of a variable X is:

$$E[X] = \sum_x x \, P(X=x)$$

More generally, for a function of X:

$$E[g(x)] = \sum_x g(x) P(x)$$ common example: $g(x) = x^2$

Conditional Expectation: $E[Y|X=x] = \sum_Y y \, P(Y=y|X=x)$

$E[X]$ is useful for making "guesses" of X's value, b/c $f=E[X]$ is the function which minimizes the <u>expected square error</u> $E[(f-X)^2]$

and $E[Y|X=x]$ is a best guess of Y given we have observed $X=x$, b/c $g=E[Y|X=x]$ minimizes $E[(f-Y)^2 | X=x]$

(This assumes implicitly that $P(X)$ or $P(Y|X=x)$ is approximately symmetric. If these distributions are skewed, it's better to use other statistics, such as the <u>median</u>, which minimizes the expected <u>absolute</u> error.)

$\Rightarrow$ Guessing and error measures are closely related to <u>loss functions</u> in ML/SL and optimization problems more generally

statistics con't

- Variance – measure how much a numeric quantity varies around it's expected value
- Covariance – measure how much two numeric quantities vary together.

Def: variance: $Var(X) = E[(X-\bar{X})^2]$, $\bar{X} = E[X]$

let's simplify this → $= E[(X - E[X])^2]$

$$= E[X^2 - 2E[X]X + E[X]^2]$$

b/c E[·] is linear (see next page) → $= E[X^2] - 2E[X]E[X] + E[E[X]^2]$

$$= E[X^2] - 2E[X]^2 + E[X]^2$$

$E[const] = const$

$$= E[X^2] - E[X]^2$$

Def covariance: $Cov(X,Y) = E[(X-\bar{X})(Y-\bar{Y})]$

notice that: $Var(X) = Cov(X,X)$
$Cov(X,Y) = 0$ if X, Y are uncorrelated

Standard deviation: $\sigma_x = \sqrt{Var(x)}$, sometimes write $Var(x) = \sigma_x^2$

(S.D. of X is nice b/c it has the same units as x itself

Properties of mean and variance

1. Expectation is a linear function

- $E[c_1 X + c_2 Y] = c_1 E[X] + c_2 E[Y]$, c_1, c_2 consts.
- $E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E[X_i]$ in general

2. Variance obeys:

- $Var(X) \geq 0$ w/ equality iff random variable is a constant
- $Var(X + c) = Var(X)$
- $Var(cX) = c^2 Var(X)$
- $$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j Cov(X_i, X_j)$$
$$= \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i \neq j} c_i c_j Cov(X_i, X_j)$$

- If R.V.s X_i are uncorrelated:

$$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i)$$ b/c $Cov(X_i, X_j) = 0$ if uncorr. & $i \neq j$