Week 14: Primer on probability, likelihood, & Bayesian methods

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

1

Announcements

- Lecture will span Monday & Wednesday
- Leftover time on Wed. will be for the tutorial (labeled as "Week 15")
- No homework this week
- Class presentations on **Dec. 2nd**
- No lab on Dec. 4th
- Final paper due on **Dec. 9th at 10pm**

2

Statistics vignette

- How often should 40-year-olds have a mammogram to screen for breast cancer?
- In 2009, US gov't advised 40-year-olds **NOT** to have annual mammograms (caused an uproar)
- **WHY???**

https://www.hopkinsmedicine.org/news/media/releases/despite_new_recommendations_women_in_40s_continue_to_get_routine_mammograms_at_same_rate

3

Some relevant numbers

- Mammograms catch breast cancer in 40-year-olds 80% of the time (true positive rate) (National Cancer Institute)
- False positive rate is 10% (*New England Journal of Medicine*)

What's the probability an asymptomatic person w/ no history of breast cancer has it, given an abnormal mammogram?

4

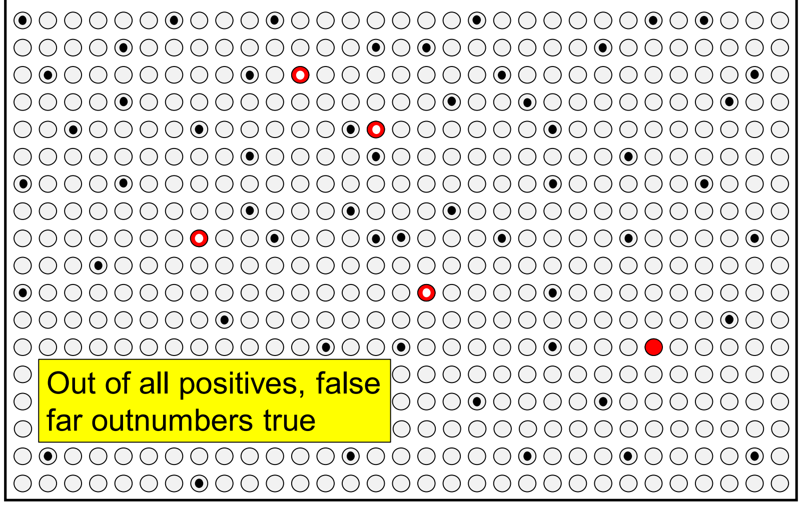
QUITE low!

- The answer is **3%**
- This is because the background rate of breast cancer is very low: 0.4% (*Cancer*, *Journal of the American Medical Association*)

https://www.komen.org/breast-cancer/screening/when-to-screen/average-risk-women/
https://www.breastcancer.org/research-news/screening-at-40-instead-of-50-saves-lives

A visual depiction

Healthy
False positive
Cancer
True positive



Or one can use Bayes' theorem

- $P(C|+) = \frac{P(+|C)P(C)}{P(+)}$
- $P(+|C)$ = true positive rate = 0.8
- $P(C)$ = background cancer rate = 0.004
- $P(+)$ = positive mammogram rate = (true positives + false positives) / everyone = 0.1
- $P(C|+) = \frac{0.8 \times 0.004}{0.1} = 0.03$

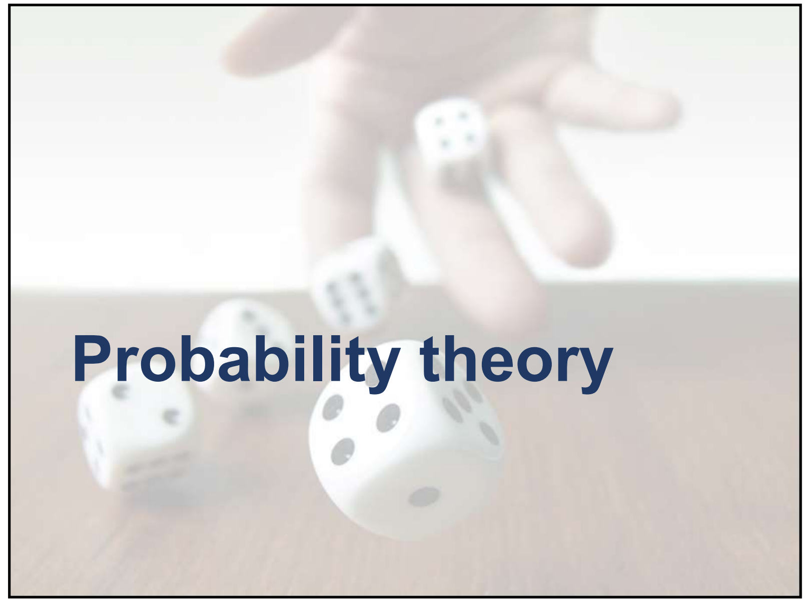
Lecture outline

- Probability theory
  - Fundamentals of probability
  - Probability distributions
- Likelihood
  - Fundamentals of likelihood & maximum likelihood estimation
  - Hypothesis testing & model selection
- Bayesian
  - Subjective probability
  - Prior information & calculating the posterior

Foundations of model building

- I covered a **LOT** of methods in this course, so you can pick the best one for your question
- Even better is constructing **your own** method or model, **perfectly** suited for your question
- The topics covered in this lecture are the foundation for building your own models

9



Probability theory

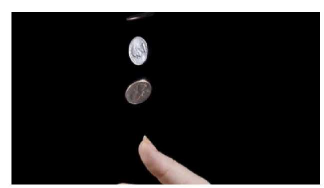
10

Fundamentals of probability

11

What is probability?

- **How likely an outcome is**
  - E.g., What is the probability a coin flip is heads?
- Can never predict outcome w/ 100% certainty because of variation in process of interest
- Probability lies at the foundation of all statistics



12

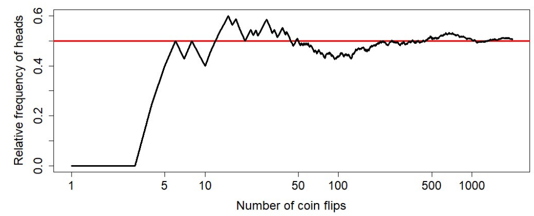
The frequentist perspective

- *P = # outcomes / # trials* (range: 0–1)
- Specifically, the relative frequency of some outcome as # trials → infinity
- E.g., how would you infer probability of rolling a 1?

13

The frequentist perspective

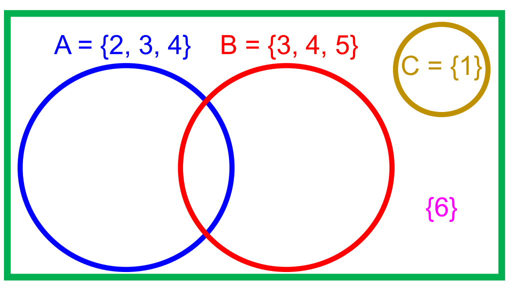
- SUPER empirical! Requires that a trial can be repeated many, many times (at least in principle)
- Infinity trials not feasible, so this is considered theoretically or need representative sample
  - E.g., statistician John Kerrich flipped coin 2,000 times while imprisoned by Nazis in WW2



14

Probability definitions

- **Sample space**: set of all possible outcomes
- **Event**: any subset of the sample space
  - E.g., P(A) = probability of event A happening



15

First axiom of probability

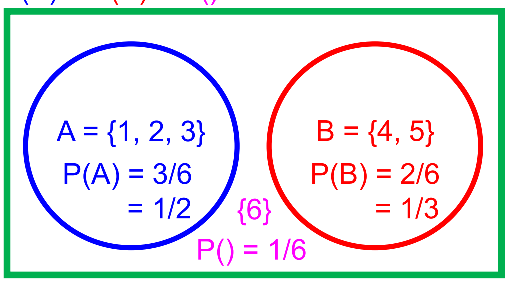
- **The sum of all probabilities of outcomes within sample space = 1.0**
  1. Events must be **mutually exclusive**: no elements in common, e.g., {1, 2} and {3, 4}
  2. Events must be **exhaustive**: covers all possible outcomes, e.g., {1, 2, 3} and {4, 5, 6}

16

First axiom of probability

- Outcomes are <u>mutually exclusive</u> (no overlap) & <u>exhaustive</u> (together, comprise entire sample space: {1, 2, 3, 4, 5, 6})

P(A) + P(B) + P() = 3/6 + 2/6 + 1/6 = 1.0



17

Complements

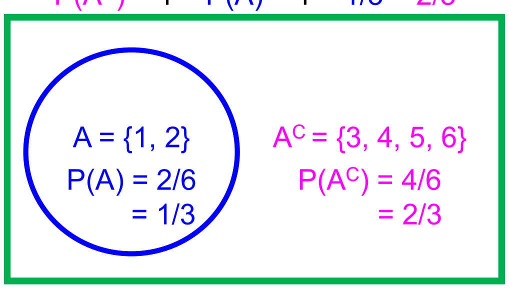
- Can use 1st axiom to calculate probability of the <u>**complement**</u> of an event, i.e., the probability an event *doesn't* happen in sample space
- Complements are represented with a ′ or ^{c}
  - $P(A') = P(A^c) = 1 - P(A)$
  - E.g., $P(\{1\}^c) = 1 - P(\{1\}) = 1 - 1/6 = 5/6$
- Works because complements are *always* mutually exclusive and exhaustive

18

Complements

- By definition, are <u>mutually exclusive</u> (no overlap) & <u>exhaustive</u> (together, comprise entire sample space)

$P(A^C) = 1 - P(A) = 1 - 1/3 = 2/3$



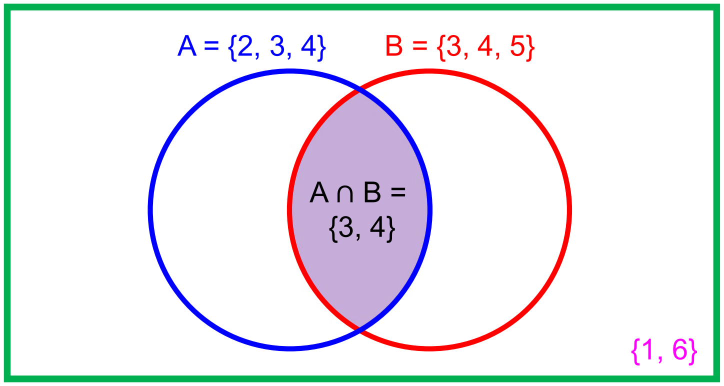
19

Intersections

- The common outcomes between two (or more) events
- Intersections are represented with ∩
- "AND" statement in logic; & in R
- E.g., {1, 2, 3} ∩ {3, 4, 5} = {3}
- P({1, 2, 3} ∩ {3, 4, 5}) = P({3}) = 1/6

20

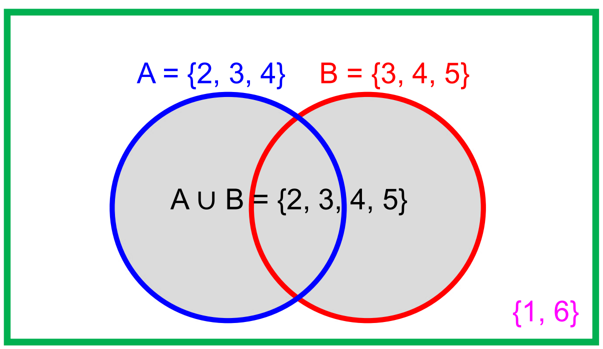
Intersections



Unions

- The union of two (or more) events is the set of all outcomes that are in either or both events
- Unions are represented with a ∪
- “OR” statement in logic; | in R
- E.g., {1, 2, 3} ∪ {3, 4, 5} = {1, 2, 3, 4, 5}
- P({1, 2, 3} ∪ {3, 4, 5}) = P({1, 2, 3, 4, 5}) = 5/6

Unions

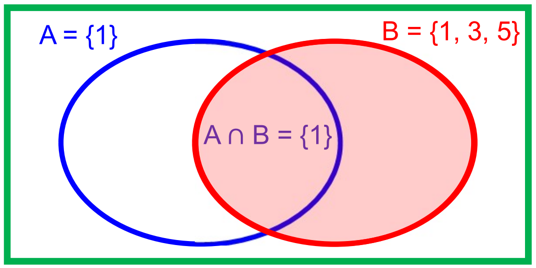


Conditional probability

- Probability of an event, given prior occurrence of another event
- P(A | B) is probability that A happens, given that B has happened
  - “Probability of A given B” or “probability of A conditional on B”
- E.g., probability of rolling a one, given that you rolled an odd number

Conditional probability

- P(A | B), e.g., P({1} | {1, 3, 5})
- B becomes new sample space
- W/in B, calculate probability of A also happening

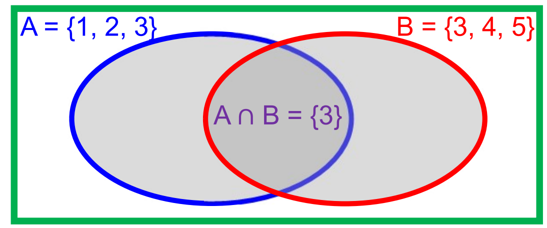


$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(\{1\}|\{1,3,5\}) = \frac{1/6}{3/6} = 1/3$$

The addition rule

- Associated with unions, e.g., {1, 2, 3} ∪ {3, 4, 5}
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B)$ **<u>IF</u>** A and B are mutually exclusive, i.e., $P(A \cap B) = 0$



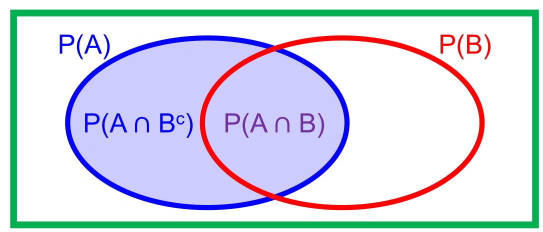
$$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}$$

The multiplication rule

- Associated with intersections
- $P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$
  - Just a rearrangement of conditional probability formula
- If A happening does not affect P(B) and vice versa, A and B are **<u>independent</u>** events
  - $P(B|A) = P(B)$ and $P(A|B) = P(A)$
- **<u>IF</u>** A & B are independ., $P(A \cap B) = P(A) \times P(B)$
  - E.g., if relative frequency of dominant allele in population is p, relative frequency of homozygous dominant genotype is p^2

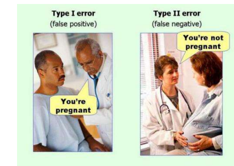
Law of total probability

- Transforms conditional and/or intersection probabilities into **<u>marginal probability</u>** AKA **<u>unconditional probability</u>** (e.g., P(A), P(B))
- P(A) = P(A ∩ B) + P(A ∩ B^C)
- Also P(A) = P(B) x P(A | B) + P(B^C) x P(A | B^C)



Questions?

E.g., Type I error

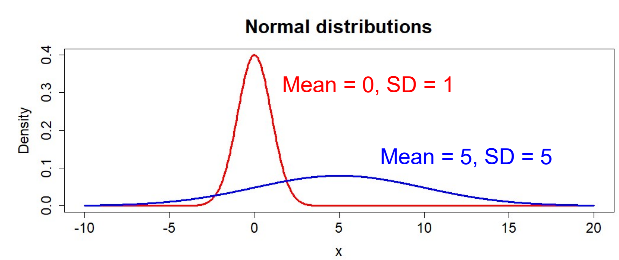


- What is the probability of getting **<u>at least</u>** one false positive, given 100 tests and $\alpha = 0.05$?
- This is the complement of getting **<u>no</u>** false positives for all 100 tests
  1. P(false pos.) = 0.05
  2. P(false pos.^{c}) = 1 – 0.05 = 0.95
  3. [P(false pos.^{c})]^{100} = $0.95^{100} \approx 0.006$
     - Assumes tests are independent
  4. 1 – [P(false pos.^{c})]^{100} ≈ 1 – 0.006 ≈ **<u>0.994</u>**
  5. $\mathbf{1-(1-\alpha)^n}$

Probability distributions

What is a probability distribution?

- A function that describes how likely certain values are in a **<u>random variable</u>**, i.e., where outcomes are not 100% predictable
- Shape is described by **<u>parameters</u>**



Two types of distributions

1. **<u>Discrete probability distributions</u>**
   - Describes random variables whose outcomes are finite or countable (e.g., integers)
   - AKA probability **<u>mass</u>** function (PMF)
   - E.g., binomial distribution, Poisson distribution
2. **<u>Continuous probability distributions</u>**
   - Describes random variables whose outcomes can take on any value within a smooth interval
   - AKA probability **<u>density</u>** function (PDF)
   - E.g., normal distribution, lognormal distribution

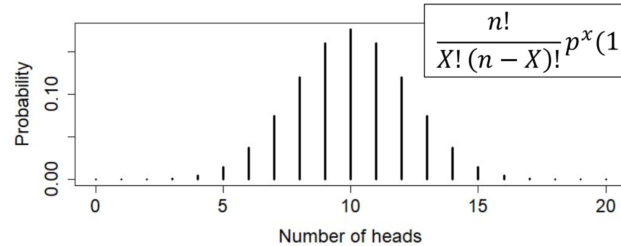
33

*parameters in red

<u>Discrete</u>: binomial PMF

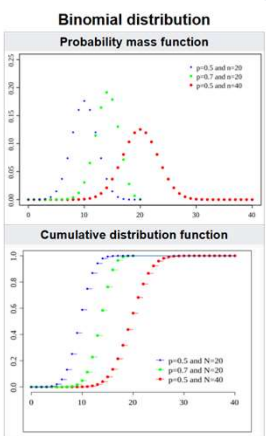
- Describes probability of getting X successes in n trials, given a probability of success, p
- E.g., probability of flipping X heads in 20 flips, given probability of heads is 0.5

$$\frac{n!}{X!\,(n-X)!}p^{x}(1-p)^{n-X}$$



34

Wikipedia pages are great for probability distributions!



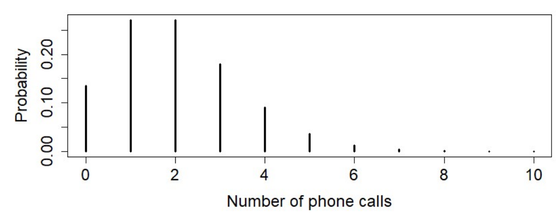
Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial $q = 1-p$
Support	$k \in \{0,1,\ldots,n\}$ – number of successes
PMF	$\binom{n}{k}p^k q^{n-k}$
CDF	$I_q(n-\lfloor k\rfloor, 1+\lfloor k\rfloor)$ (the regularized incomplete beta function)
Mean	np
Median	$\lfloor np\rfloor$ or $\lceil np\rceil$
Mode	$\lfloor (n+1)p\rfloor$ or $\lceil (n+1)p\rceil - 1$
Variance	$npq = np(1-p)$
Skewness	$\frac{q-p}{\sqrt{npq}}$
Excess kurtosis	$\frac{1-6pq}{npq}$

35

*parameters in red

<u>Discrete</u>: Poisson PMF

- Describes probability of getting X occurrences of an event in a fixed area or time period, given an average # occurrences in said area/time (λ)
- E.g., probability of getting X phone calls in an hour, given average calls/hour is 2

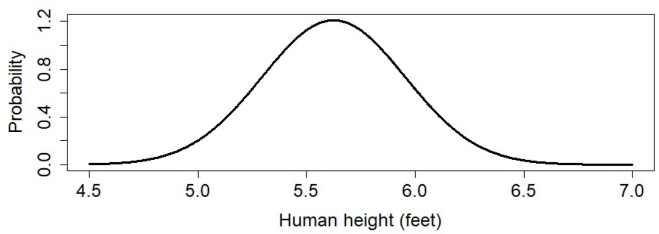


36

*parameters in red

Continuous: normal PDF

- Describes how likely a value of X is, given the mean (μ) and SD (σ)
- X is outcome of additive processes
- E.g., human heights in a population

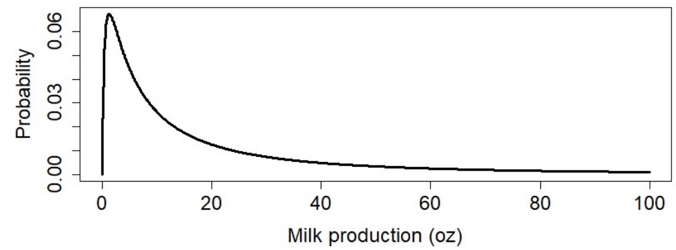


37

*parameters in red

Continuous: lognormal PDF

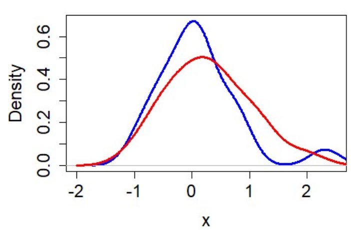
- Describes how likely a value of X is, given the mean (μ) and SD (σ) in log space
- X is product of multiplicative processes
- E.g., milk production by cows

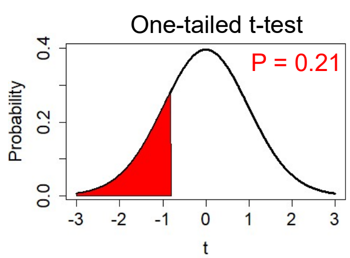


38

Cumulative distribution function (CDF)

- Calculates probability that X is ≤ some value for **any** distribution
- E.g., P-values: probability of getting null statistic more extreme than observed statistic, assuming H_0 is true

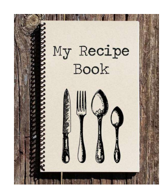




39

Simplified modeling recipe

1. Figure out the P(A) or P(A | B) that addresses your research question & how to derive it from other probabilities, e.g., P(B | A), P(B)
2. Figure out how to represent each probability with a distribution
3. Carry out the math to get your probability model



40

An example from paleoanthro.

SCIENCE ADVANCES | RESEARCH ARTICLE

ANTHROPOLOGY

Temporal evidence shows *Australopithecus sediba* is unlikely to be the ancestor of *Homo*

2019

Andrew Du* and Zeresenay Alemseged

1. Figure out probability:

$$P(X_A > X_D) = \int_{t=0}^{\infty} P(X_A > t)P(X_D = t)dt \quad (3b)$$

2. Represent w/ distributions:

terval). Using the exponential cumulative distribution function, this probability is

$$P(X_A > t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t} \quad (3c)$$

The second probability in the integral ($P[X_D = t]$; Eq. 3b) can be calculated using the exponential probability density function (i.e., the probability that X_D takes on some value, t), so

$$P(X_D = t) = \lambda e^{-\lambda t} \quad (3d)$$

3. Carry out the math:

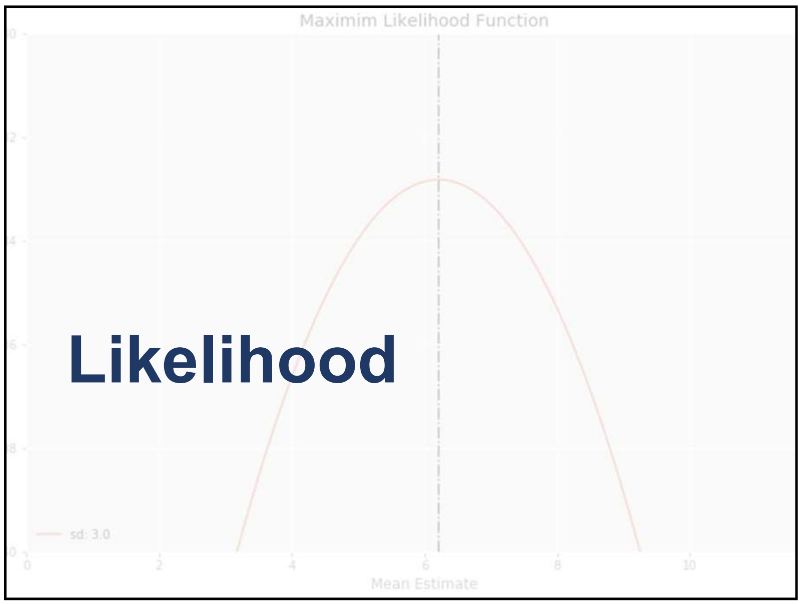
Substituting Eqs. 3c and 3d into Eq. 3b, we get

$$P(X_A > X_D) = \int_{t=0}^{\infty} e^{-\lambda t}\lambda e^{-\lambda t}dt \quad (3e)$$

41

Questions?

42



Likelihood

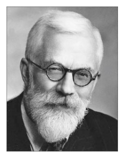
43

Fundamentals of likelihood & maximum likelihood estimation

44

What is likelihood?

- A principled **frequentist** framework for statistical inference and modeling
  1. Parameter estimation
  2. Hypothesis testing
  3. Model selection
- A lot of methods can be derived, and thus unified, with likelihood (e.g., t-tests, OLS)
- First developed by R.A. Fisher



45

Defining likelihood

- Traditional frequentist tests calculate P(Data | Model), e.g., P-value
- Likelihood inverts the conditional probability to get L(Model | Data)
- E.g., probability asks what is the probability of getting 4 heads out of 10 coin flips (data), given that p = 0.5 (assumed model parameter)
- Likelihood asks how likely is p = 0.5 (model parameter), given that you get 4 heads out of 10 coin flips (data)

46

Another example

Weather (parameter)

Attire (data)	Cold	Warm
Jacket	0.8	0.1
T-shirt	0.2	0.9
Total	**1.0**	**1.0**

- P(attire | cold)
- L(weather | jacket)
- Likelihoods don't have to sum to one (not true probabilities)
- P(jacket | cold) = L(cold | jacket)

- Probability is a statement about observed data
- Likelihood is a statement about the parameter(s)

From Wang (2010)

47

Maximum likelihood estimate (MLE)

- Likelihood provides a framework for estimating unknown parameter(s) in a system
- MLE is the parameter value(s) that makes the observed data most probable (i.e., has the highest likelihood)
  - E.g., on previous slide, MLE is "cold"
  - Given the person wore a jacket, "cold" has a higher likelihood (0.8) than "warm" (0.1)

48

E.g., lion stalking success

- In Ngorongoro Crater (Tanzania), Elliott et al. (1977) found that lions had 34 out 157 successful stalks of wildebeest and zebra
- **<u>Of the entire (partially sampled) population of lions, what is the rate of successful stalks at Ngorongoro?</u>**

MLE of stalking success rate

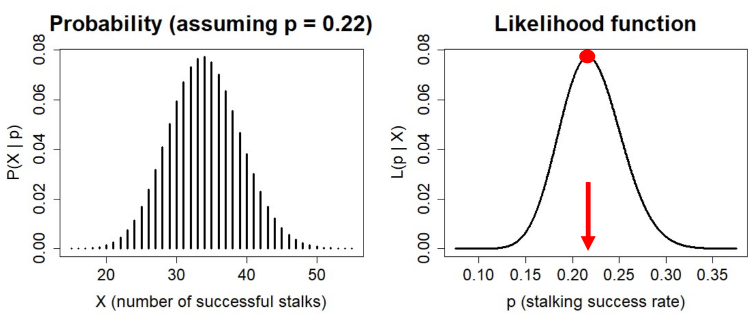
X: number of successful stalks (34)

n: number of total stalks (157)

p: stalking success rate

- Assuming stalks are independent, can model rate (p) with binomial distribution
- A good naïve guess of p is 34 / 157 ≈ 0.22 (cf. frequentist definition of probability)

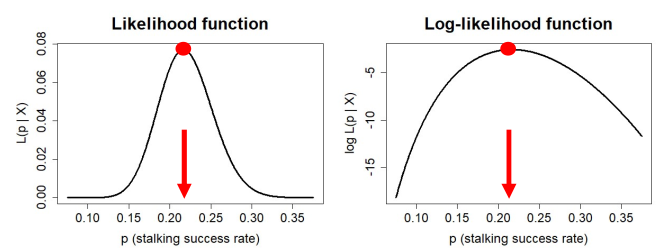
MLE of stalking success rate



- **<u>Likelihood function</u>** – likelihood values as fxn. of parameter values (numerically equal to P(X | p))
- MLE = value of p that gives the highest likelihood

MLE of stalking success rate

- How to find MLE of p (tip of bell curve)?
- Take derivative of function, set it to zero (maxima of function), and solve for p
- Easier to do this with **<u>log-likelihood</u>** function



Maximum at same value of p

Getting MLE of p

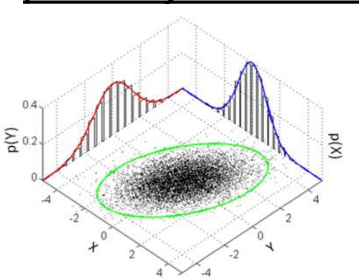
1. $L(p|X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$ (binomial dist.)
2. $\log L(P|X) = \log\left(\frac{n!}{X!(n-X)!}\right) + X\log(p) + (n-X)\log(1-p)$
3. $\frac{d}{dp}\log L(p) = 0 + \frac{X}{p} - \frac{n-X}{1-p}$
4. $\frac{X}{\hat{p}} - \frac{n-X}{1-\hat{p}} = 0$ (the hat indicates an estimated parameter)
5. $\boldsymbol{\hat{p} = \frac{X}{n}}$
6. $\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (derived using log-likelihood function)

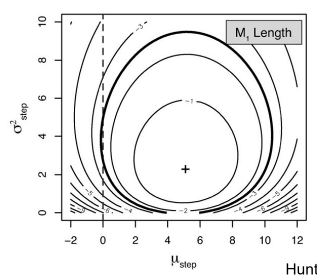
MLE of stalking success rate

- So $\hat{p} = \frac{X}{n} = \frac{34}{157} \approx 0.22$
- This matches our naïve estimate, but we derived it formally
- MLE has good statistical properties: as $n \to \infty$,
  - Estimate is unbiased
  - Has the smallest possible variance among all unbiased estimators
  - Sampling distribution is normal

MLE with >1 parameters

- Find MLE of parameters simultaneously using log-likelihood function
- Likelihood function constructed from **joint probability distribution**





Hunt, 2006

Questions?

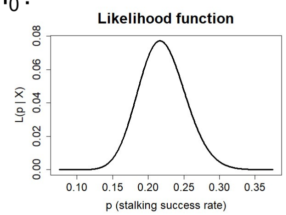
Hypothesis testing

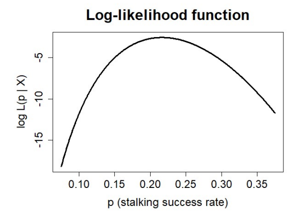
- Let's say the literature says lion stalking success rates should be 0.1 (H_0 → M_0)
- We can test if the underlying parameter generating our observed data (i.e., MLE; M_1) is significantly different from 0.1
- <u>**M_0**</u>: P(success. stalk) = p = 0.1
  - zero free parameters
- <u>**M_1**</u>: p is free to vary (i.e., is estimated)
  - one free parameter
- Which model fits the data better?

57

Hypothesis testing

- $L(M_0 \mid X)$ = 7.7e-6; log $L(M_0 \mid X)$ = -11.8
- $L(M_1 \mid X)$ = 0.08; log $L(M_1 \mid X)$ = -2.6
- More complex models (more free parameters) always fit the data better
- How to know if M_1 fits data significantly better than M_0?





58

Likelihood ratio tests (LRT)

- If models are nested (complex model has ≥1 extra parameter), can use LRT to test if more complex model (M_1) is sig. better than simpler one (M_0)

$$LR = -2(\log L[M_0] - \log L[M_1])$$

- If M_0 is supported by the data, the two likelihoods should not differ by more than sampling error
- LR follows χ^2 dist. w/ degrees of freedom = difference in # free parameters between models
- For our example, $LR = -2(-11.8 - [-2.6]) = 18.4$
  - P = 1.8e-5

59

Model selection

Hirotugu Akaike

- Formalized way of competing hypotheses (models) against each other on equal footing
- The Akaike Information Criterion (AIC) balances goodness of fit (logL) and model complexity (K = # free parameters)

$$\text{AIC} = -2\log L + 2K$$

- AIC measures amount of information lost in approximating reality w/ model (lower AIC is better)
- Can transform into weights (sum to one across models, w/ larger weights → more support)

60

Our lion stalking example

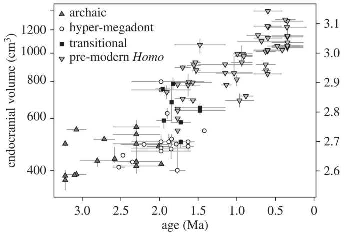
Model	Descrip.	logL	K	AIC*	AIC weight
M_0	p = 0.1	-11.8	0	23.55	2.7e-4
M_1	p free to vary	-2.6	1	7.13	0.9997

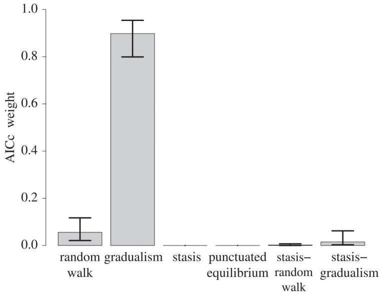
*General rule: >2 difference in AIC → good support

61

Another example

- How did hominin brain size increase over time?





Du et al. 2018

62

Questions?

63

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

Bayesian paradigm

64

Bayesian paradigm

- There are two main differences compared to the frequentist paradigm:

1. A different definition of probability (i.e., **<u>subjective probability</u>**)
2. The incorporation of **<u>prior information</u>** in models

65

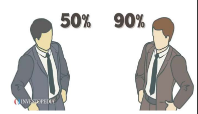
Probability: a test

- Are you a frequentist or Bayesian?
- I flip a coin and then cover it with my hand

- What is the probability that the coin is heads?

66

Subjective probability



- Recall that the frequentist definition of probability is the relative frequency of some outcome as # trials → infinity
- Problematic for unique events
  - E.g., what is the probability that Vermont is larger than New Hampshire?
- Subjective probability quantifies our **<u>uncertainty</u>** or **<u>degree of belief</u>** in some event, whether it is repeatable or not

67

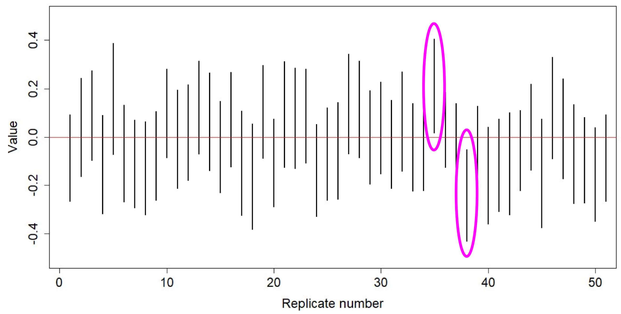
Treatment of parameters

- The <u>frequentist</u> paradigm treats parameters as **<u>fixed</u>** quantities
  - E.g., the average height of <u>*all*</u> humans on Earth is equal to one (unknown) value
- Randomness is introduced by the sampling process (e.g., each sample of data gives a different mean estimate of height)
- The <u>Bayesian</u> paradigm treats parameters themselves as **<u>random b/c of our uncertainty about them</u>**

68

E.g., 95% CIs

- **Frequentist**: fixed parameter is either inside CI or not (in long run, 5% of CIs exclude parameter)



69

E.g., 95% CIs

- **Frequentist**: fixed parameter is either inside CI or not (in long run, 5% of CIs exclude parameter)
- **Bayesian**: treats parameter as random due to uncertainty, so 95% CI interpreted as a 95% probability parameter is inside CI
  - A Bayesian confidence interval is called a **credible interval**

70

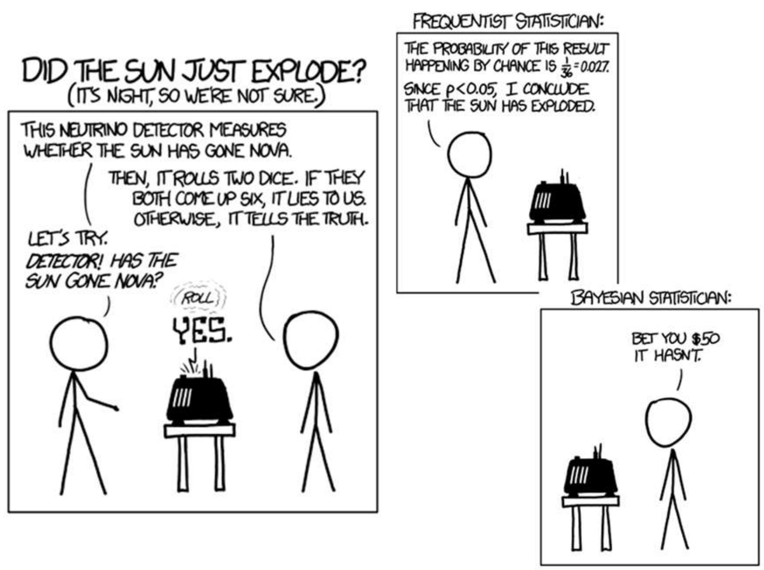
Questions?

71

Prior information

- The main debate: should one incorporate prior knowledge about quantity of interest, external to the dataset?
- E.g., I carry out a presidential approval rating poll in an area and got 41%, even though other organizations got around 55%
- Should I adjust my results upwards (e.g., average 41% and 55%)?
- Unscientific and unethical? Or smart to "stand on the shoulder of giants"?

72



73

Prior information

- It can be argued that **every** researcher incorporates their own biases into their studies
- E.g., researcher finds implausible results and runs experiment for longer
- The Bayesian paradigm enables researchers to incorporate prior information in their models in a principled, formalized, & transparent manner

74

How to incorporate prior information?


Thomas Bayes

- Bayes' theorem (or Bayes' rule):

Prior probability: quantifies prior information

Likelihood: summarizes the data

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

Posterior probability: probability of hypothesis/model given the data

Scaling factor: Makes area under $P(H|D)$ distribution sum to one

75

76

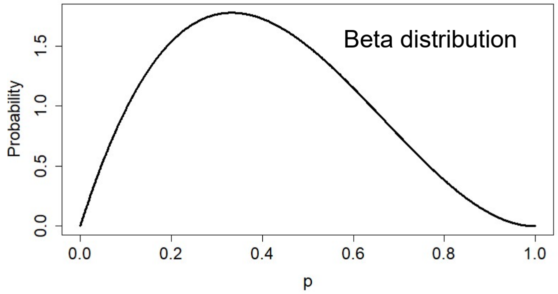
Example: lost wallets

- What is the probability (p) that police officers will return lost wallets to owners but steal some money?

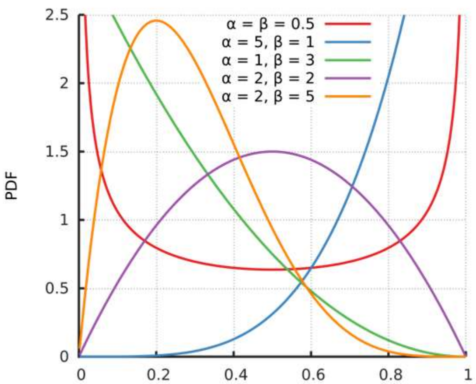
From Wang (2010)

1. Formulate the prior, P(H)

- What do **<u>YOU</u>** think the probability is?
- Quantify this with a probability distribution

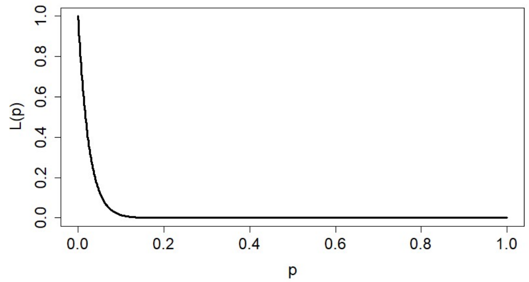


The beta distribution is very flexible



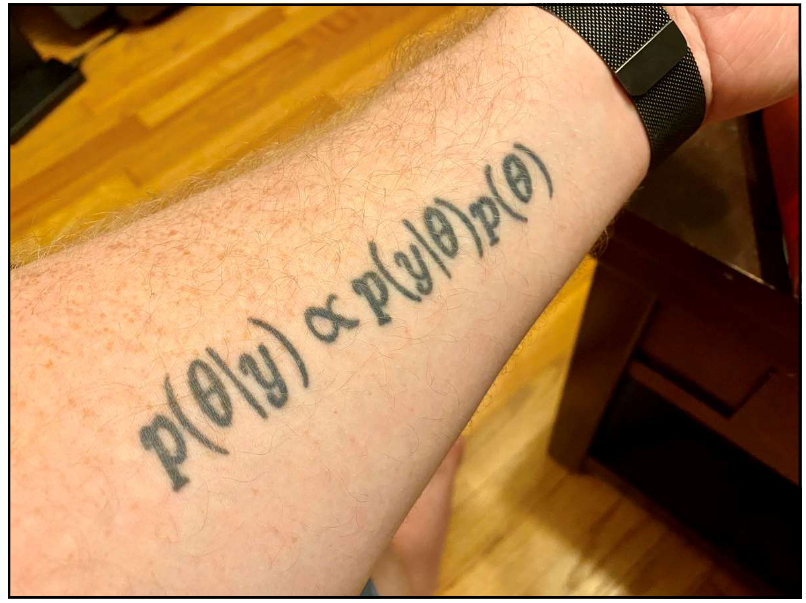
2. Collect data & formulate likelihood function, P(D | H)

- In experiment run by *Primetime*, 40 out of 40 officers returned wallets w/ **<u>NO</u>** money missing
- Create likelihood function using binomial distrib.



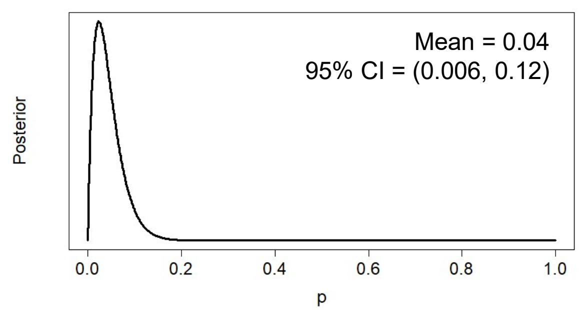
3. Update prior w/ data, P(H | D)

- Bayes' theorem: $P(H|D) = \frac{P(H)P(D|H)}{P(D)}$
- Oftentimes, we ignore the scaling factor, $P(D)$
  - Probability distribution looks identical; only scale of y-axis changes
  - OK b/c only care about which values of p are more probable relative to each other
- So $P(H|D) \propto P(H)P(D|H)$
  - Prior can be thought of as weighting certain values of p in the likelihood
  - If prior is uninformative (all values of p likely), just doing a likelihood analysis

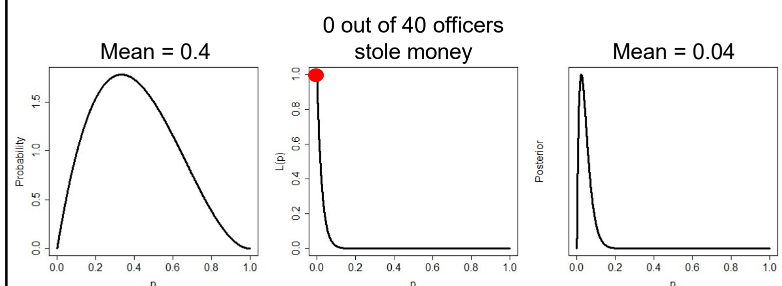


3. Update prior w/ data, P(H | D)

- $P(H|D) \propto P(H)P(D|H)$
- Multiply prior by the likelihood



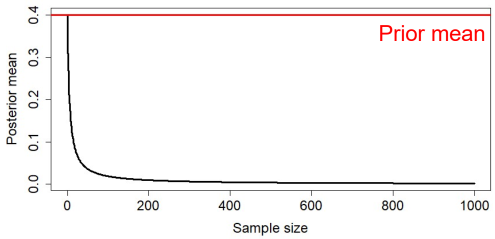
Prior, likelihood, posterior



- **<u>NB</u>**: w/ zero officers stealing money, MLE of p is zero
- Do we actually expect <u>NO</u> officers to steal money?

Influence of dataset size

- Assuming we collect more data and still no officers steal money
- At small n, prior dominates
- At larger n, MLE dominates ("data speak for themselves")



Questions?

Summary

- Probability theory quantifies how likely certain events are and how likely certain values in data are
- Likelihood is a principled framework for inferring parameters, testing hypotheses, and comparing models
- Bayesian methods offer a formalized framework for combining prior information w/ the likelihood

E.g., birthday paradox

- What is the probability **at least** two out of three people will share the same birthday?
- This is the complement of **nobody** sharing a birthday
  - P(1st birthday on any day) = 365 / 365 = 1
  - P(2nd birthday not on that day) = 364 / 365
  - P(3rd birthday not on either day) = 363 / 365
  - P(no shared birthdays) = $1 \times \frac{364}{365} \times \frac{363}{365} \approx 0.99$
  - P(no shared birthdays^c) = 1 – 0.99 ≈ **0.008**