Probability Theory and Mathematical Statistics

Jingcheng Liu

Goals

- A quick introduction to the mathematics behind statistics
- Understand basic terminology
- Know how to formulate a statistical problem

What is statistics

Microsoft Bing

研究表明

WEB IMAGES VIDEOS ACADEMIC DICT MAPS ⋮ MORE

About 27,800,000 results Any time

研究人员招募了36名健康成年人，并将其随机分配到发酵或高纤维饮食方案小组中，使其维持该方案10周，并在实验开展前3周、采取分组饮食后10周，以及结束实验饮食方案4周后，采集参与者血液和粪便样本进行分析。 **研究人员发现**，这两种饮食方式对肠道微生物和免疫系统产生了不同影响。 食用酸奶、发酵白干酪、泡菜和其他发酵蔬菜及相关饮品等，会增加人体微生物多样性，食用量越大，影响越强。 “这是一个惊人的**发现**。 ”斯坦福大学微生物学和免疫学副教授Justin Sonnenburg说，该**研究**说明了简单的饮食改变是如何重塑健康人体内的微生物群的。

新研究表明发酵食品益处多----中国科学院

www.cas.cn/kj/202107/t20210714_4798455.shtml

Was this helpful?

研究表明：2050年全球8.4亿多人腰痛，女性病例高于男性 ...

https://www.thepaper.cn/newsDetail_forward_23233269

Web May 26, 2023 · 一项基于30多年数据的分析表明，全球腰痛病例数量正在增加。. 模型显示，到2050年，由于人口增长和人口老龄化，将有8.43亿人受到这种疾病的影响。. 相关论文将发表于6月刊的《柳叶刀-风湿病学》。. 由于腰痛是全球人类致残的主要原因，研究人员 ...

Bai du 百度

研究表明 百度一下

网页 资讯 贴吧 知道 文库 图片 地图 采购 视频 更多

百度为您找到相关资讯58个 按焦点排序 全部资讯

澳学者研究发现年龄差距大的夫妻彼此更满意,美学者意见正好相反...

1小时前 一项研究表明,年龄相差数十岁的夫妻,对爱情的满意度可能会更高。 据《每日邮报》5月31日报道,澳大利亚迪肯大学的一个科学小组认为,与年龄相仿的情侣相比,有巨大年龄差的夫妇表示,他们更加信任彼此,嫉妒心也更少。 研究人员称:“超过四...

荆楚网

谁更爱咿咿呀呀?研究表明一岁以内男宝宝“话”更多

5小时前 然而,依据国际学术期刊《交叉科学》5月31日刊载的一篇研究论文,一岁以内的男婴比女婴更爱咿咿呀呀,连论文作者也对此感到意外。他们猜测,这可能是人类进化使然。 2020年12月 10日,在位于加沙城的一家联合国近东巴勒斯坦难民救济和工程处...

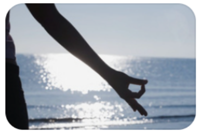
大洋网

中国科学家揭示早期地球海洋维持漫长缺氧原因

9小时前 据了解,很多研究表明,前寒武纪海洋在很大程度上是以缺氧分层为主,氧化可能仅存在于海洋的表层浅水等区域。但由于人们缺乏能够直接追踪古海洋溶解磷含量的定量指标,因而无法准确定量古海洋中溶解磷的时空波动。 李超教授团队在2021年研发了能...

央视网

最新研究:定期冥想可有效避免负面情绪_新闻频道_中国青年网



13小时前 据《瑞士资讯》30日报道,最新研究表明,定期的正念冥想可有效避免负面情绪。 报道称,苏黎世联邦理工学院的研究人员将261名志愿者随机分为两组,进行了为期两周的观察实验:一组人闭眼打坐,...

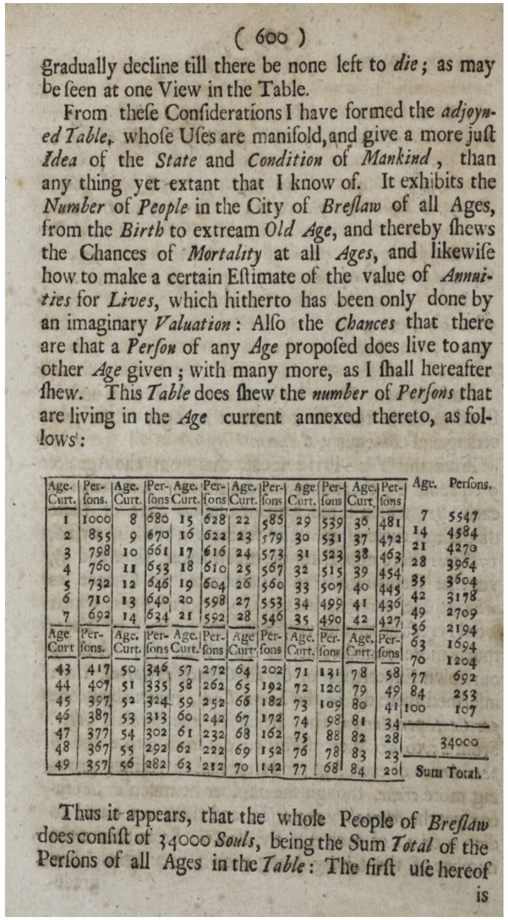
中国青年网

The tale of Edmond Halley’s life table

Published in 1693, Halley found applications of his life table in:
- Estimate the proportion of men in a population that could bear arms
- Pricing life annuity
- …

Data Summary
- Many details/information are being thrown away:
  - How/when/where are they collected
- Abstraction/summary/modelling: to generalize
  - “To think is to forget a difference, to generalize, to abstract.”
    -- *Funes the Memorious* by Jorge Luis Borges

Statistical modelling by probability (stochastic modelling)
- How do we quantify the quality of a model?
- How confident are we that a pattern is real?



(600)

gradually decline till there be none left to *die*; as may be ſeen at one View in the Table.

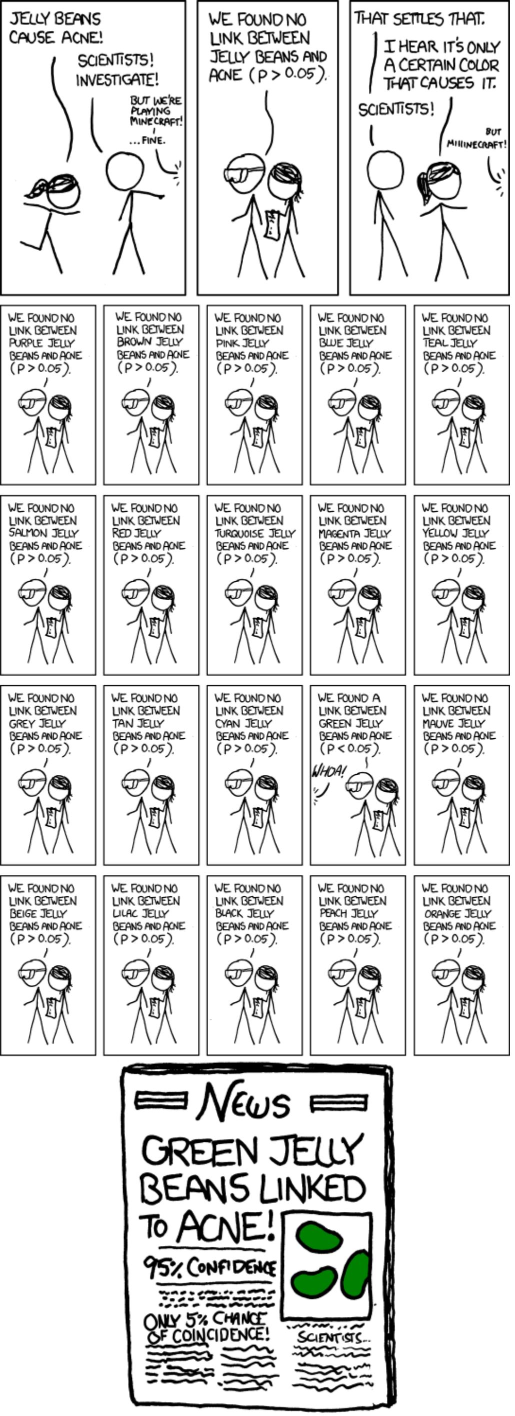
From theſe Conſiderations I have formed the *adjoyned Table*, whoſe Uſes are manifold, and give a more juſt *Idea* of the *State* and *Condition* of *Mankind*, than any thing yet extant that I know of. It exhibits the *Number* of *People* in the City of *Breſlaw* of all Ages, from the *Birth* to extream *Old Age*, and thereby ſhews the Chances of *Mortality* at all *Ages*, and likewiſe how to make a certain Eſtimate of the value of *Annuities* for *Lives*, which hitherto has been only done by an imaginary *Valuation*: Alſo the *Chances* that there are that a *Perſon* of any *Age* propoſed does live to any other *Age* given; with many more, as I ſhall hereafter ſhew. This *Table* does ſhew the *number* of *Perſons* that are living in the *Age* current annexed thereto, as follows:

Age. Curt.	Perſons.	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons
1	1000	8	680	15	628	22	585	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age Curt	Perſons.	Age. Curt.	Perſons	Age. Curt.	Perſons	Age Curt.	Perſons	Age. Curt.	Perſons	Age. Curt.	Perſons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Perſons.
7	5547
14	4584
21	4270
28	3964
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

Thus it appears, that the whole People of *Breſlaw* does conſiſt of 34000 *Souls*, being the Sum *Total* of the Perſons of all Ages in the *Table*: The firſt uſe hereof is

Source: royalsocietypublishing.org



Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Article | **Authors** | **Metrics** | **Comments** | **Media Coverage**

Correction

25 Aug 2022: Ioannidis JPA (2022) Correction: Why Most Published Research Findings Are False. PLOS Medicine 19(8): e1004085. https://doi.org/10.1371/journal.pmed.1004085 | **View correction**

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

See:
https://xkcd.com/882/

Note: there is even a talk show lamenting about "p-hacking"

The problems with food questionnaires go even deeper. They aren't just unreliable, they also produce huge data sets with many, many variables. The resulting cornucopia of possible variable combinations makes it easy to p-hack your way to sexy (and false) results, as we learned when we invited readers to take an FFQ and answer a few other questions about themselves. We ended up with 54 complete responses and then looked for associations — much as researchers look for links between foods and dreaded diseases. It was silly easy to find them.

Our shocking new study finds that ...

EATING OR DRINKING	IS LINKED TO	P-VALUE
Raw tomatoes	Judaism	<0.0001
Egg rolls	Dog ownership	<0.0001
Energy drinks	Smoking	<0.0001
Potato chips	Higher score on SAT math vs. verbal	0.0001
Soda	Weird rash in the past year	0.0002
Shellfish	Right-handedness	0.0002
Lemonade	Belief that "Crash" deserved to win best picture	0.0004
Fried/breaded fish	Democratic Party affiliation	0.0007
Beer	Frequent smoking	0.0013
Coffee	Cat ownership	0.0016
Table salt	Positive relationship with Internet service provider	0.0014
Steak with fat trimmed	Lack of belief in a god	0.0030
Iced tea	Belief that "Crash" didn't deserve to win best picture	0.0043
Bananas	Higher score on SAT verbal vs. math	0.0073
Cabbage	Innie bellybutton	0.0097

SOURCE: FFQ & FIVETHIRTYEIGHT SUPPLEMENT

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mistaken. Researchers typically express the confidence in their data in terms of p-value: the probability that a perceived result is actually the result of random variation. The value of p (for "probability") is a way of measuring the extent to which a data set provides

a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed nonspecifically as an investigation of possible associations between party affiliation and mathematical reasoning across contexts. The null hypothesis is that the political context is irrelevant to the task, and the alternative hypothesis is that context matters and the dif-

This *multiple comparisons* issue is well known in statistics and has been called "p-hacking" in an influential 2011 paper by the psychology researchers Joseph Simmons, Leif Nelson, and Uri Simonsohn. Our main point in the present article is that it is possible to have multiple potential comparisons (that is, a data analysis whose details are highly contingent on data, invalidating published p-val-

Science Isn’t Broken

It’s just a hell of a lot harder than we give it credit for.

By Christie Aschwanden
Graphics by Ritchie King
Filed under Scientific Method

Hack Your Way To Scientific Glory

You’re a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office**. Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you’ll need to prove that they are “statistically significant” by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [ ] Governors
- [x] Senators
- [ ] Representatives

How do you want to measure economic performance?

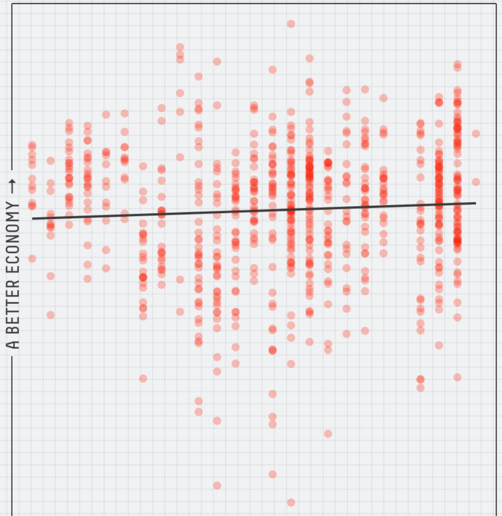
- [ ] Employment
- [x] Inflation
- [ ] GDP
- [x] Stock prices

Other options

- [ ] Factor in power
  Weight more powerful positions more heavily
- [ ] Exclude recessions
  Don’t include economic recessions

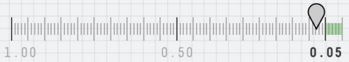
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Almost

Your **0.08** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

https://fivethirtyeight.com/features/science-isnt-broken/

Sally Clark's case

Sally Clark was convicted for murdering her two sons, when both died within weeks after birth
Her conviction was largely based on a mis-use of statistics, for ruling out sudden infant death syndrome

- Recall the "Dominating false positive" example during probability lectures

$$\Pr[\text{a rare natural event} \mid \text{innocence}] \neq \Pr[\text{innocence} \mid \text{a rare natural event}]$$

See also: https://en.wikipedia.org/wiki/Sally_Clark

- https://en.wikipedia.org/wiki/Base_rate_fallacy
- https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy
- TED talk by Peter Donnelly: How stats fool juries

Statistical questions: more examples

- Travel insurance: Should you purchase insurance for your next flight?
  - The same flight has a delay record of 53%
  - The insurance starts paying whenever the flight is delayed for more than 10 minutes

- Clinical trial:
  - Treatment I: “100% effective”, cured 3 out of 3.
  - Treatment II: “95% effective”, cured 19 out of 20.
  - Treatment III: “90% effective”, cured 90 out of 100.

  - Which treatment is more effective?

- Dam construction in hydrology:
  - Dam should be high enough *for most floods*
  - Should not be unnecessarily high (expensive)

- Should you allow AdBlocker on your website?
- Why museums charge differently based on group?
  - What’s the basis of student discount?
- Frequency analysis in cryptography
  - Deciphering the Enigma in World War II

What is common in these questions?

- In expectation
- Need to quantify chance (Is it worth it? Is it effective?)
- Significance of our conclusion

Probability vs. Statistics

In probability, we often consider a well-defined/idealized random experiment.

- Flip a fair/unbiased coin
- Roll a fair/unloaded dice
- Draw a card

Probability vs. Statistics

All models are wrong
but some are useful

In statistics, we first need a (probabilistic) model of the real world.
Randomness can come from:

- the probabilistic model (biased coin, flight delay)
- using “simple process”+ “noise” in the modelling

A statistic is anything that can be computed from collected data.
The goal is often to make inferences from collected data.

Statistical mechanics, but not probabilistic mechanics;
Probabilistic combinatorics, but not statistical combinatorics
(not to confuse with combinatorial statistics)

Probability vs. Statistics

In probability: Compute probabilities from a parametric model with known parameters

Previous studies found the treatment is 80% effective. Then we expect that for a study of 100 patients, on average 80 will be cured. And the probability that at least 65 will be cured is at least 99.99%.

In statistics: Estimate the probability of parameters given a parametric model and collected data from it

Observe that 78/100 patients were cured. We will be able to conclude that: if we repeat this experiment, then we are 95% confident that the number of cured patients are between 69 to 87.

Later in class: can be derived from Chernoff-Hoeffding bound

A toy model

Say we model the problem of predicting flight delays as
independent Bernoulli's with unknown parameter p

<u>Why probabilistic modelling?</u>

We abstract our "lack of knowledge" about the physical laws of flight delays, using stochasticity.

<u>Why Bernoulli?</u>

We assume that the problem follows a distribution that conceptualizes what is a typical instance:

If we see a new flight, how much delay do we expect to see?

A toy model

Say we model the problem of predicting flight delays as independent Bernoulli's with unknown parameter p

We observe 100 times.

Given that there were 55 delays, what is a good estimate for p ?

How about $\hat{p} = 0.55$?

In general, a statistical model is a ***parametric*** probabilistic model

Maximum likelihood estimates (MLE)

MLE asks:

Which parameter maximizes the chances of seeing the observed data?

This is known as a point estimate.

Compare with: outputting an interval, or an estimated p.d.f.

In our toy model of independent Bernoulli's with unknown parameter p

$$\Pr[55\ heads \mid p] = \binom{100}{55} p^{55}(1-p)^{45}$$

Likelihood, or likelihood function

Maximum likelihood estimates (MLE)

MLE asks:

Which parameter maximizes the chances of seeing the observed data?

In our toy model of independent Bernoulli's with unknown parameter p

$$\Pr[55\ heads \mid p] = \binom{100}{55} p^{55}(1-p)^{45}$$

$$\frac{d}{dp}\Pr[55\ heads \mid p] = \binom{100}{55}\left(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}\right)$$

Setting derivative to 0 we have $\hat{p} = 0.55$

Equivalently, one can try to maximize log-likelihood

Maximum likelihood estimates (MLE)

MLE = sample mean holds for

- n independent Bernoulli's with unknown parameter p
- Poisson with unknown parameter
- Gaussian

(derivations are similar)

Algorithms for MLE: often iterative, see Expectation-Maximization algorithm

Maximum likelihood estimates (MLE)

Many real-world applications:

Lifetime of a light bulb, or your hard disk: often modelled by an exponential distribution with unknown parameter

Reliability	Temp.	Operating	0°C to 70°C (Measured by S.M.A.R.T. Temperature Proper airflow recommended)	
		Non-Operating	-40°C to 85°C	
	Humidity		5% to 95% non-condensing	
	Shock	Non-Operating	1,500G(Gravity), duration: 0.5ms, 3 axis	
	Vibration	Non-Operating	20~2,000Hz, 20G	
	MTBF		1.5 million hours	
Warranty[6)]	TBW		600TB	1,200TB
	Period		5 years limited	

> 171 years!

Mark and recapture method for estimating the size of a population: recall balls and bins experiments

Bayesian inference

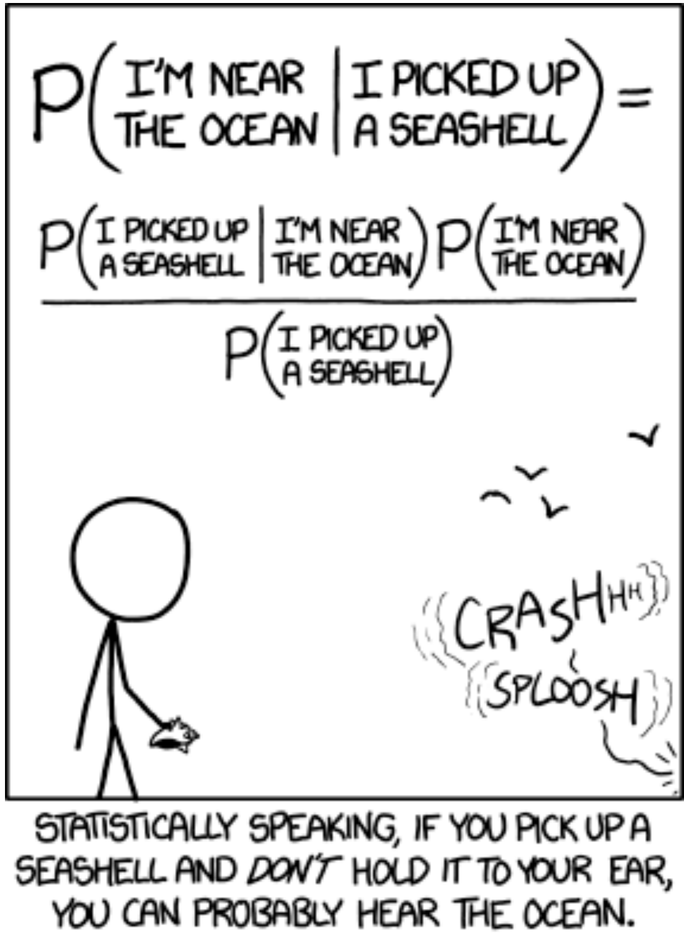
We associate a prior distribution to the unknown model and parameters

Then we apply Bayes' law to transfer this from the collected data to a distribution on the unknown parameters.

This is called the posterior distribution.

Types of problems:

- Estimation
- Hypothesis testing



Maximum A Posteriori (MAP)

We are estimating p given data

Why maximize $\Pr[\text{data}|\,p]$ instead of $\Pr[p|\text{data}]$?

Recall Bayes' law:

Posterior likelihood Prior

$$\Pr[p|\text{data}] = \frac{\Pr[\text{data}|\,p]\,\Pr[p]}{\Pr[\text{data}]}$$

Need to choose a prior, and different priors lead to different estimate

Example: IMDB score

Estimation theory

We saw two estimators for the parameter p given n iid samples from $Bernoulli(p)$:

- MLE:
  - Frequentists approach
  - Inference based on likelihood
  - p is an unknown parameter, we estimate it purely based on data

Parameter: fixed
Data: random

- MAP:
  - Bayesian approach
  - p is unknown, but it follows a prior distribution
  - Inference based on posterior distribution
  - we estimate it based on the observed data and our prior belief

Parameter: random
Data: fixed

- How do we compare different estimators?
  - Bayesian: mean squared error
  - Frequentist: risk

Minimum mean squared error estimators

Mean squared error: in our toy model, if p is random and $\hat{p}$ is a constant

$$\mathbb{E}(\hat{p} - p)^2$$

Observe that $\mathbb{E}(\hat{p} - p)^2 = var(p) + (\mathbb{E}p - \hat{p})^2$ is minimized when

$$\hat{p} \coloneqq \mathbb{E}p$$

If $\hat{p}$ depends on the data, the mean squared error is then:

$$\mathbb{E}[(\hat{p} - p)^2 | data]$$

By a similar argument, MMSE is given by $\hat{p} \coloneqq \mathbb{E}[p | data]$

Compared with Bayesian MMSE: expectation is taken over prior!

Frequentists risk

Consider n iid samples from $Bernoulli(p)$ with an unknown parameter p:

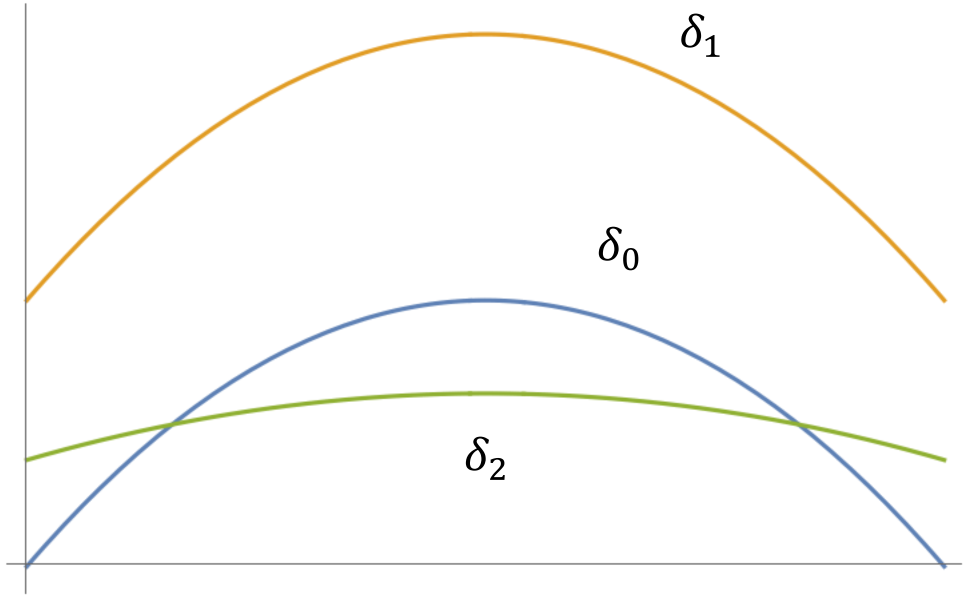
- Loss: $L(p, \delta)$ measures how bad an estimate is
  - $L(p, \delta) = (p - \delta)^2$ is known as the squared loss
- Risk of an estimator:
  - Expected loss, where expectation is taken over the distribution of data

Example

- $\delta_0(X_1, X_2, \dots, X_n) = \sum_i \frac{X_i}{n}$
- $\mathbb{E}\delta_0(X_1, X_2, \dots, X_n) = p$, so unbiased
- Risk under mean squared loss: $\mathbb{E}(p - \delta_0)^2 = Var(\delta_0) = \frac{p(1-p)}{n}$

Consider two other estimators: $\delta_1 = \frac{1+\sum_i X_i}{n}$, $\delta_2 = \frac{5+\sum_i X_i}{10+n}$

Let's plot their risk functions



Compared with Bayesian MMSE: expectation is taken over prior!

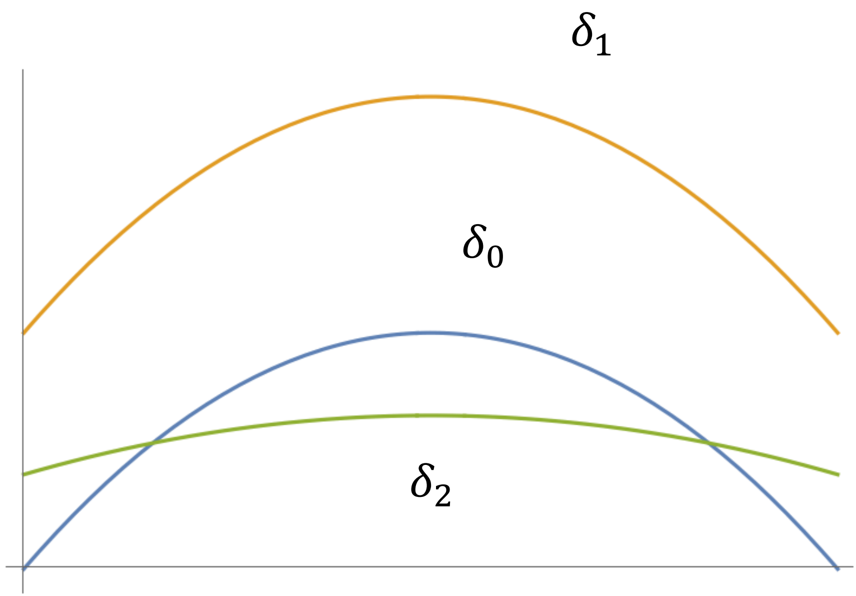
Frequentists risk

Example

- $\delta_0(X_1, X_2, \ldots, X_n) = \sum_i \frac{X_i}{n}$
- $\mathbb{E}\delta_0(X_1, X_2, \ldots, X_n) = p,$ so unbiased
- Risk under mean squared loss: $\mathbb{E}(p - \delta_0)^2 = Var(\delta_0) = \frac{p(1-p)}{n}$

Consider two other estimators: $\delta_1 = \frac{1+\sum_i X_i}{n}$, $\delta_2 = \frac{5+\sum_i X_i}{10+n}$

δ_1 may look stupid. But δ_0 vs δ_2 is trickier…



Rules for choosing THE BEST one:

- Average risk: choose a prior over p $\rightarrow$ Bayesian!
- Worst-case risk: minimax estimator
- Only consider unbiased estimator: (see next)

Sufficient statistics

Suppose $X_1, \dots, X_n \sim Bernoulli(p)$:

Consider $T(X) \coloneqq X_1 + \cdots + X_n \sim Bin(n, p)$

$X_1, \dots, X_n \;\rightarrow\; T(X)$ can throw away information

To estimate p however, $T(X)$ is just as informative as $X_1, \dots, X_n$

$$\Pr[X = x | T = t] = \frac{\Pr[X = x, T = t]}{\Pr[T = t]}$$

<u>**Definition**</u>. $T(X)$ is a <u>***sufficient statistic***</u> for a parameter p, if the distribution of X does not depend on p given T

Sufficient statistics are the only information needed to build an estimator

Parameter $\Rightarrow$ Sufficient statistics $\Rightarrow$ Samples

Minimal sufficiency

There are many sufficient statistics for our toy model:

- $X_1, \dots, X_n$
- $X_{\sigma(1)}, \dots, X_{\sigma(n)}$
- $X_1 + \cdots + X_n$

Definition. $T(X)$ is a ***minimal sufficient statistic*** for a parameter p, if T is sufficient, and any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f

Intuitively, minimal sufficient statistics are the most efficient statistics capturing all the information about the parameter

Roughly speaking, if T determines the likelihood ratio in a “one-to-one fashion”, then T is minimal sufficient.

See also: Fisher’s factorization theorem.

Sufficiency principle: Rao-Blackwellization

Let $T(X)$ be a sufficient statistic, and $\delta_0(X)$ an estimator.
Consider a new estimator $\delta_1(T(X)) \coloneqq \mathbb{E}[\delta_0(X) \mid T(X)]$

For convex losses, the Rao–Blackwell estimator δ_1 is at least as good as δ_0

In practice, can lead to enormous difference.

See Textbook [BT] page 426 Exercises for examples

Minimum variance unbiased estimator (optional)

Lehmann–Scheffé theorem roughly says that any unbiased estimator through a *complete* and sufficient statistic, is the ***unique*** minimum variance unbiased estimator.

Complete statistic
Roughly, T is complete if there is no non-trivial estimate of 0 through T
Different estimates of T lead to different distributions

See also: Cramér–Rao bound, which gives a bound on how efficient an unbiased estimator can be.

Caution about unbiasedness (optional topic)

Not always a good idea to insist unbiasedness, because Cramér–Rao bound may not be achievable

Example:

Data samples $X \sim Bin(1000, p)$, want to estimate $\Pr[X \geq 500]$.

One can show that the minimum variance unbiased estimator is just $\mathbb{I}[X \geq 500]$

- This means that if $X = 500$, our estimate is 1
- if $X = 499$, our estimate is 0

Confidence interval

How do you interpret the results of an estimation?

- By LLN/CLT, any (asymptotically) unbiased estimator converges to the true parameter as the sample size tends to infinity
- By Chernoff-Hoeffding bound, we also get a finite size bound

Suppose $X_1, \dots, X_n \sim Bernoulli(p)$ are iid r.v. , and $S_n = \sum_i X_i$ then for any $t > 0$

$$\Pr[|S_n - np| \geq t] \leq 2\mathrm{e}^{-\frac{2\mathrm{t}^2}{\mathrm{n}}}$$

Setting $\alpha = 2\mathrm{e}^{-\frac{2\mathrm{t}^2}{\mathrm{n}}}$, we have $t = \sqrt{\frac{n \ \ln(2/\alpha)}{2}}$.

This means that with probability $1 - \alpha$,

$$p \in \left(\frac{S_n}{n} - \sqrt{\frac{\ln\left(\frac{2}{\alpha}\right)}{2n}}, \quad \frac{S_n}{n} + \sqrt{\frac{\ln(2/\alpha)}{2n}} \right).$$

It is important to note that this probability is **over the distribution of S_n**

Confidence interval: interpretations

A 95% confidence interval is NOT an interval that contains the true parameter with probability at least 95%

The confidence interval is a function of the data

After observing the data, the confidence interval is a fixed interval

It either contains the true parameter, or not

To bring back probabilistic interpretation:

- Consider repeating the experiments, over and over again
  - Now you have new, fresh, random data, so that the confidence interval can be treated as a random object over *future repeated experiments* of the assumed statistical/generative model
  - In particle physics, usually a five-sigma rule, unless ground-breaking discovery
- Bayesian approach: credible region
  - Only way to conclude from what we have already observed

Recall Probability vs. Statistics

In probability: **Compute probabilities from a parametric model with known parameters**

Previous studies found the treatment is 80% effective. Then we expect that for a study of 100 patients, on average 80 will be cured. And the probability that at least 65 will be cured is at least 99.99%.

In statistics: **Estimate the probability of parameters given a parametric model and collected data from it**

Observe that 78/100 patients were cured. We will be able to conclude that: **if we repeat this experiment**, then we are 95% confident that the number of cured patients are between 69 to 87.

Bayesian vs. frequentist

Bayesian

- Inference based on posterior
- A feature or a bug: Prior
- Probabilities can be interpreted
- Prior is made explicit
- Prior can be subjective
- No canonical prior: can change under re-parameterization
- Hierarchical Bayesian, graphical model
- Computation/sampling of posterior can be hard
  - Frontiers of many research

Frequentist

- Inference based on likelihood
- No prior
- Objective – everyone gets the same answer
- Often gets mis-interpreted
- Needs to completely specify an experiment AND the data analysis, before collecting data and actually doing the analysis
- No adaptive re-use of the same dataset
  - There is an entire field for systematically coping with adaptive data analysis