18.650 – Fundamentals of Statistics

1. Introduction and probability

Goals

Goals:

- To give you a solid introduction to the mathematical theory behind statistical methods;
- To provide theoretical guarantees for the statistical methods that you may use for certain applications.

At the end of this class, you will be able to

1. From a real-life situation, formulate a statistical problem in mathematical terms
2. Select appropriate statistical methods for your problem
3. Understand the implications and limitations of various methods

Why statistics?

In the press

The New York Times

THE UPSHOT

Nike Says Its $250 Running Shoes Will Make You Run Much Faster. What if That's Actually True?

An analysis of nearly 500,000 running times estimates the effect of shoes on race performance.

https://www.nytimes.com/interactive/2018/07/18/upshot/nike-vaporfly-shoe-strava.html
Citation/Attribution: Article (c) New York Times

MIT Technology Review

Data Mining Reveals the Way Humans Evaluate Each Other

Vast databases of soccer statistics expose the limited way human observers rate performance and suggest how they can do significantly better.

Object Source/ URL: https://www.technologyreview.com/s/609760/data-mining-reveals-the-way-humans-evaluate-each-other/
Citation/Attribution -- Article from the MIT Technology Review. (c) MIT

In businesses

Harvard Business Review

How Vineyard Vines Uses Analytics to Win Over Customers

TECHNOLOGY DIGITAL ARTICLE by Dave Sutton

A case study on how personalization is changing retail.

SAVE SHARE JUNE 08, 2018

ttps://hbr.org/2018/06/how-vineyard-vines-uses-analytics-to-win-over-customers
Citation/Attribution -- Article and Image Copyright © 2019 Harvard Business School Publishing. All rights reserved

FAST COMPANY

AppNexus is key to AT&T's plans to use HBO for more consumer data

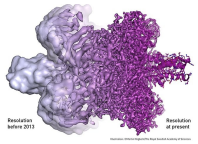
New WarnerMedia CEO John Stankey says HBO is going to "change direction a little bit," and it's all about the advertising.

ttps://www.fastcompany.com/90188017/appnexus-is-key-to-atts-plans-to-use-hbo-for-more-consumer-data
Citation/Attribution -- Article by Jeff Beer on Fast Company & Inc © 2019 Mansueto Ventures,

In science and engineering

The Guardian

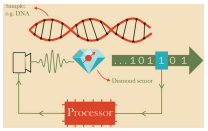
What is cryo-electron microscopy, the Nobel prize-winning technique?



Object Source / URL*
https://www.theguardian.com/science/2017/oct/04/what-is-cryo-electron-microscopy-the-chemistry-nobel-prize-winning-technique
Citation/Attribution -- Image (c) The Guardian

IEEE SPECTRUM

Measuring Tiny Magnetic Fields With an Intelligent Quantum Sensor



https://spectrum.ieee.org/tech-talk/biomedical/devices/measuring-tiny-magnetic-fields-with-an-intelligent-quantum-sensor
Citation/Attribution -- Article Image (c) International Journal of Electrical, Electronics and Data Communication.

On TV

LAST WEEK TONIGHT WITH JOHN OLIVER

"Last Week Tonight with John Oliver": Scientific Studies

Object Source / URL*
https://www.youtube.com/watch?v=0Rnq1NpHdmw&has_verified=1 Citation/Attribution Photo of John Oliver © 2019 Home Box Office, Inc. All Rights Reserved

Data Science and the Art of Producing Entertainment at Netflix

Object Source / URL*
https://medium.com/netflix-techblog/studio-production-data-science-646ee2cc21a1 Citation/Attribution Image on the Medium website (c) Netflix corporation.

Statistics, Data Science . . . and all that

Statistics, Data Science, Machine Learning, Artificial Intelligence

What's the difference?

Statistics, Data Science . . . and all that

Statistics, Data Science, Machine Learning, Artificial Intelligence

What's the difference?

- All use data to gather insight and ultimately make decisions
- Statistics is at the core of the data processing part
- Nowadays, computational aspects play an important role as data becomes larger

Computational and statistical aspects of data science

- Computational view: data is a (large) sequence of numbers that needs to be processed by a relatively fast algorithm: approximate nearest neighbors, low dimensional embeddings, spectral methods, distributed optimization, etc.
- Statistical view: data comes from a **random process**. The goal is to learn how this process works in order to make predictions or to understand what plays a role in it.

To understand randomness, we need PROBABILITY.

Probability

- Probability studies randomness (hence the prerequisite)
- Sometimes, the physical process is completely known: dice, cards, roulette, fair coins, . . .

Rolling 1 die:

- Alice gets \$1 if # of dots ≤ 3
- Bob gets \$2 if # of dots ≤ 2

$E[A] = \frac{1}{2} \cdot \$1 = \$.5$

$E[B] = \frac{1}{3} \cdot \$2 = \$.6$

Who do you want to be: Alice or Bob?

Rolling 2 dice:

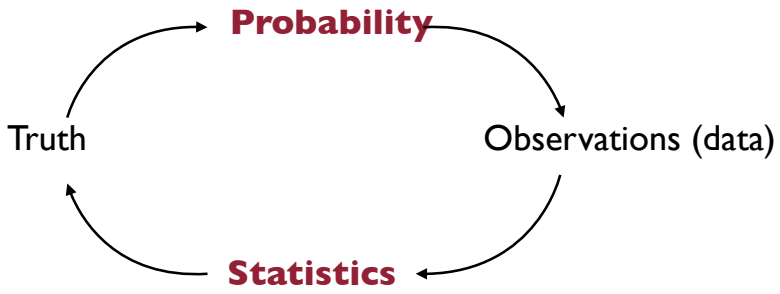
- Choose a number between 2 and 12
- Win \$100 if you chose the sum of the 2 dice

Which number do you choose?

Statistics and modeling

- Dice are well known random process from physics: 1/6 chance of each side (no need for data!), dice are independent. We can deduce the probability of outcomes, and expected \$ amounts. This is **probability**.
- How about more complicated processes? Need to estimate parameters from data. This is **statistics**
- Sometimes real randomness (random student, biased coin, measurement error, . . .)
- Sometimes deterministic but too complex phenomenon: **statistical modeling**

  Complicated process "=" Simple process + random noise
- (good) Modeling consists in choosing (plausible) simple process **and** noise distribution.



Statistics vs. probability

Probability Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with 99.99% chances.

Statistics Observe that 78/100 patients were cured. We (will be able to) conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients

What this course is about

- Understand **mathematics** behind statistical methods
- Justify quantitive statements given modeling assumptions
- Describe interesting mathematics arising in statistics
- Provide a math toolbox to extend to other models.

What this course is **not** about

- Statistical thinking/modeling (e.g., 15.075)
- Implementation (e.g. IDS.012)
- Laundry list of methods (e.g. AP stats)

What this course is about

- Understand **mathematics** behind statistical methods
- Justify quantitive statements given modeling assumptions
- Describe interesting mathematics arising in statistics
- Provide a math toolbox to extend to other models.

What this course is **not** about

- Statistical thinking/modeling (e.g., 15.075)
- Implementation (e.g. IDS.012)
- Laundry list of methods (e.g. AP stats)

Let's do some statistics

The kiss

Le baiser. Auguste Rodin. 1882.

Object Source / URL*
http://www.musee-rodin.fr/en/collections/sculptures/kiss Citation/Attribution Photo (c) Musée Rodin

The kiss

Le baiser. Auguste Rodin. 1882.

Object Source / URL*
http://www.musee-rodin.fr/en/collections/sculptures/kiss Citation/Attribution Photo (c) Musée Rodin

The kiss

Object Source / URL*
https://www.nature.com/articles/421711a Citation/Attribution – SpringerNature.com

Statistical experiment

"A neonatal right-side preference makes a surprising romantic reappearance later in life."

- Let p denote the proportion of couples that turn their head to the right when kissing.
- Let us design a statistical experiment and analyze its outcome.
- Observe n kissing couples times and collect the value of each outcome (say 1 for RIGHT and 0 for LEFT);
- *Estimate* p with the proportion $\hat{p}$ of RIGHT.
- Study: "Human behaviour: Adult persistence of head-turning asymmetry" (Nature, 2003): $n = 124$ and 80 to the right so

$$\hat{p} = \frac{80}{124} = 64.5\,\%$$

Random intuition

Back to the data:

- 64.5% is much larger than 50% so there seems to be a preference for turning right.
- What if our data was RIGHT, RIGHT, LEFT ($n = 3$). That's 66.7% to the right. Even better?
- Intuitively, we need a large enough sample size n to make a call. How large?
- Another way to put the problem: for $n = 124$, what is the minimum number of couple "to the right" would you need to see to be convinced that $p > 50\%$? 63? 72? 75? 80?

We need **mathematical modeling** to understand the accuracy of this procedure?

A first estimator

Formally, this procedure consists of doing the following:

- For $i = 1, \ldots, n$, define $R_i = 1$ if the ith couple turns to the right RIGHT, $R_i = 0$ otherwise.
- The estimator of p is the sample average $\sum_{i=1}^{n} R_i = \#\{i : \ldots\}$

$$\hat{p} = \bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i.$$

What is the accuracy of this estimator ?

In order to answer this question, we propose a statistical model that describes/approximates well the experiment.

We think of the R_i's as random variables so that $\hat{p}$ is also a random variable. We need to understand its fluctuation.

Modelling assumptions

Coming up with a model consists of making assumptions on the observations $R_i, i = 1, \ldots, n$ in order to draw statistical conclusions. Here are the assumptions we make:

1. Each R_i is a random variable.
2. Each of the r.v. R_i is Bernoulli with parameter p. $R_i \sim \text{Ber}(\ldots)$
3. $R_1, \ldots, R_n$ are mutually independent.

$P(R_i = 1)$
$P(R_i = 0)$

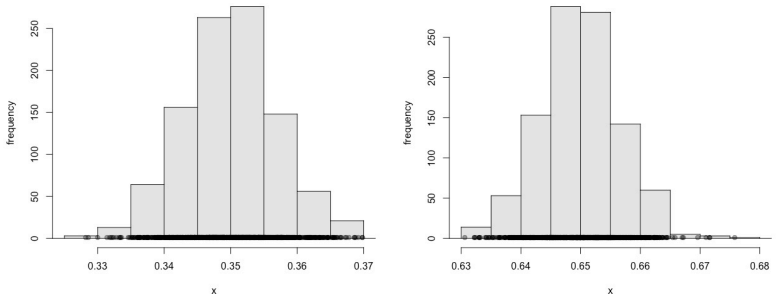
Discussion

Let us discuss these assumptions.

1. Randomness is a way of modeling lack of information; with perfect information about the conditions of kissing (including what goes on in the kissers' mind), physics or sociology would allow us to predict the outcome.
2. Hence, the R_i's are necessarily Bernoulli r.v. since $R_i \in \{0, 1\}$. They could still have a different parameter $R_i \sim \text{Ber}(p_i)$ for each couple but we don't have enough information with the data to estimate the p_i's accurately. So we simply assume that our observations come from the same process: $p_i = p$ for all i
3. Independence is reasonable (people were observed at different locations and different times).

Population vs. Samples

- Assume that there is a total **population** of 5,000 "airport-kissing" couples
- Assume for the sake of argument that $p = 35\%$ or that $p = 65\%$.
- What do **samples** of size 124 look like in each case?



Why probability?

We need to understand probabilistic aspects of the distribution of the random variable:

$$\hat{p} = \bar{R}_n = \frac{1}{n} \sum_{i=1}^{n} R_i.$$

Specifically, we need to be able to answer questions such as:

- Is the expected value of $\hat{p}$ close to the unknown p?
- Does $\hat{p}$ take values close to p with high probability?
- Is the variance of $\hat{p}$ large? I.e. does $\hat{p}$ fluctuate a lot?

We need probabilistic tools! Most of them are about **average of independent random variables**.

Probability redux

Averages of random variables: LLN & CLT

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. r.v., $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{V}[X]$.

- Laws (weak and strong) of large numbers (LLN):

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P},\ \text{a.s.}} \mu.$$

- Central limit theorem (CLT):

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

(Equivalently, $\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\sigma^2)$.)

Another useful tool: Hoeffding's inequality

What if n is not large enough to apply CLT?

Theorem (Hoeffding, 1963)
Let n be a positive integer and $X, X_1, \ldots, X_n$ be i.i.d. r.v. such that $\mu = \mathbb{E}[X]$ and

$X \in [a,b]$ almost surely ($a < b$ are given numbers)

Then,

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \varepsilon] \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}. \quad \forall \varepsilon > 0$$

This holds even for small sample sizes n.

Consequences

- The LLN's tell us that

$$\bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\ \text{a.s.}} p.$$

(what modeling assumptions did we use?)

- Hence, when the size n of the experiment becomes large, $\bar{R}_n$ is a *good* (say "*consistent*") estimator of p.

- The CLT refines this by quantifying *how good* this estimate is: for n large enough the distribution of $\hat{p}$ is almost:

$$\mathbb{P}(|\bar{R}_n - p| \geq \varepsilon) \simeq \mathbb{P}(|\mathcal{N}(0, \tfrac{p(1-p)}{n})| > \varepsilon)$$

In the Kiss example, $\mathbb{P}(|\bar{R}_n - p| \geq 0.084) \simeq 5\%$

- Hoeffding's inequality tells us that

$$\mathbb{P}(|\bar{R}_n - p| \geq 0.084) \leq 2\exp\left(-\frac{2 \cdot 124 \cdot (0.084)^2}{(1-0)^2}\right) \leq 0.35$$

The Gaussian distribution

Because of the CLT, the Gaussian (a.k.a normal) distribution is ubiquitous in statistics. It is named after German Mathematician Carl Friedrich Gauss (1777–1855) in the context of the method of *least squares* (regression).

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- $\mathbb{E}[X] = \mu$
- $\mathrm{var}(X) = \sigma^2 > 0$

Object Source / URL*
http://mathshistory.st-andrews.ac.uk/PictDisplay/Gauss.html Citation/Attribution Image from the MacTutor History of Mathematics archive (success)

Gaussian density (pdf)

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

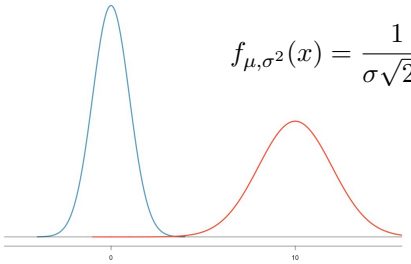


Figure 1: Two pdfs: $\mathcal{N}(0,1)$ and $\mathcal{N}(10,4)$

- Tails decay very fast (like $e^{-\frac{x^2}{2\sigma^2}}$): almost in finite interval.
- There is no closed form for their cumulative distribution function (CDF). We use tables (or computers):

$$\mathbb{P}(X \le x) = F_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Some useful properties of Gaussians

Perhaps the most useful property of the Gaussian family is that it's *invariant under affine transformation*:

- $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any $a, b \in \mathbb{R}$,

$$a \cdot X + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

- **Standardization** (a.k.a Normalization/Z-score): If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

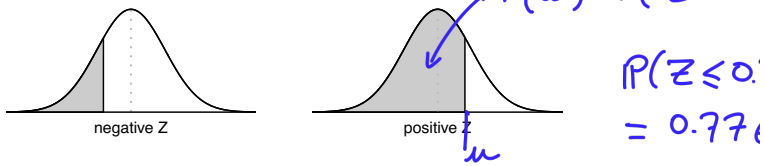
Useful to compute probabilities from CDF of $Z \sim \mathcal{N}(0,1)$:

$$\mathbb{P}(u \le X \le v) = \mathbb{P}\left(\frac{u-\mu}{\sigma} \le Z \le \frac{v-\mu}{\sigma}\right)$$

- **symmetry**: If $X \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ then $-X \sim \mathcal{N}(\mathbf{0}, \sigma^2)$: If $x > 0$

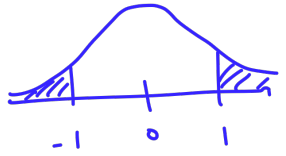
$$\mathbb{P}(|X| > x) = \mathbb{P}(X > x) + \mathbb{P}(-X > x) = 2\,\mathbb{P}(X > x)$$

Gaussian probability tables



Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Examples



Assume that $Z \sim \mathcal{N}(0,1)$ and compute

- $\mathbb{P}(Z \le 1) = 0.8413$
- $\mathbb{P}(Z \ge -1) = \mathbb{P}(-Z \le 1) = \mathbb{P}(Z \le 1) = 0.8413$
- $\mathbb{P}(|Z| > 1) = 2\,\mathbb{P}(Z > 1) = 2 \cdot (1 - 0.8413) = 0.31$

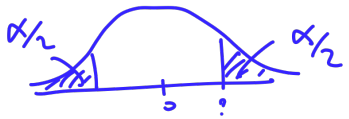
Assume that the score distribution for a final exam is approximately $X \sim \mathcal{N}(85, 4)$, compute

- $\mathbb{P}(X > 90) = \mathbb{P}\left(\frac{X-85}{2} > \frac{90-85}{2}\right) = \mathbb{P}(Z > 2.5$
- $\mathbb{P}(80 < X < 90)$

$$= \mathbb{P}(-2.5 < Z < 2.5) = 0.9876$$

More complicated: what is x such that $\mathbb{P}(X < x) = 90\%$ (90^{th} percentile?). For that we need to read the table backwards.

Quantiles



Definition

Let α in $(0,1)$. The quantile of order $1-\alpha$ of a random variable X is the number q_α such that

$$\mathbb{P}(X \leq q_\alpha) = 1-\alpha$$

Let F denote the CDF of X:

- $F(q_\alpha) =$
- If F is invertible, then $q_\alpha =$
- $\mathbb{P}(X > q_\alpha) = \alpha$
- If $X = Z \sim \mathcal{N}(0,1)$: $\mathbb{P}(|X| > q_{\alpha/2}) = \alpha$

Some important quantiles of the $Z \sim \mathcal{N}(0,1)$ are:

α	2.5%	5%	10%
q_α	1.96	1.65	1.28

We get that $\mathbb{P}(|Z| > 1.96) = 5\%$

Three types of convergence

- $(T_n)_{n\geq 1}$ is a sequence of random variables
- T is a random variable (T may be deterministic).

- Almost surely (a.s.) convergence:

$$T_n \xrightarrow[n\to\infty]{\text{a.s.}} T \quad \text{iff} \quad \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1.$$

- Convergence in probability:

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \quad \text{iff} \quad \mathbb{P}\left[|T_n - T| \geq \varepsilon\right] \xrightarrow[n\to\infty]{} 0, \quad \forall \varepsilon > 0.$$

- Convergence in distribution:

$$T_n \xrightarrow[n\to\infty]{(d)} T \quad \text{iff} \quad \mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)]$$

for all continuous and bounded function f.

Properties

- If $(T_n)_{n\geq 1}$ converges a.s., then it also converges in probability, and the two limits are equal a.s.

- If $(T_n)_{n\geq 1}$ converges in probability, then it also converges in distribution

- **Convergence in distribution** implies convergence of probabilities if the limit has a density (e.g. Gaussian):

$$T_n \xrightarrow[n\to\infty]{(d)} T \qquad \Rightarrow \qquad \mathbb{P}(a \leq T_n \leq b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \leq T \leq b)$$

Exercises

a) Is the following statement correct? "If $(T_n)_{n\geq 1}$ converges in probability, then it also converges a.s"

1. Yes
2. No

Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of r.v. such that $X_n \sim \text{Ber}(\frac{1}{n})$. Exercises b), c) and d) are about this sequence.

b) Let $0 < \epsilon < 1$, $n \geq 1$. What is the value of $P(\{|X_n| > \epsilon\})$? (answer: $\frac{1}{n}$)

c) Does $\{X_n\}$ converges in probability?

1. Yes
2. No

Exercises

d) Denote by X the limit of $\{X_n\}$ (if it exists) (that is, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$). What is the value of X?

1. X does not exist
2. 0
3. 1
4. None of the above

e) Dose $\{X_n\}$ converge in distribution?

1. Yes
2. No

f) What is the limit of the sequence $\mathbb{E}[\cos(X_n)]$ as n tends to infinity?

$\mathbb{E}[\cos(0)] = 1$

Addition, multiplication, division

. . . only for a.s. and $\mathbb{P}$. . .

Assume

$$T_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} T \qquad \text{and} \qquad U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} U$$

Then,

- $T_n + U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} T + U$,
- $T_n U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} TU$,
- If in addition, $U \neq 0$ a.s., then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \frac{T}{U}$.

⚠ In general, these rules **do not** apply to convergence (d).

Slutsky's theorem

Some partial results exist for convergence in distribution on the form of *Slutsky's theorem*.

Let $(X_n), (Y_n)$ be two sequences of r.v., such that:

$$\text{(i) } T_n \xrightarrow[n\to\infty]{(d)} T \qquad \text{and} \qquad \text{(ii) } U_n \xrightarrow[n\to\infty]{\mathbb{P}} u$$

where T is a r.v. and u is a given real number (deterministic limit: $\mathbb{P}(U = u) = 1$). Then,

- $T_n + U_n \xrightarrow[n\to\infty]{(d)} T + u$,
- $T_n U_n \xrightarrow[n\to\infty]{(d)} Tu$,
- If in addition, $u \neq 0$, then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{(d)} \frac{T}{u}$.

. . .

Taking functions

Continuous functions (for all three types) . If f is a continuous function:

$$T_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}/(d)} T \quad \Rightarrow \quad f(T_n) \xrightarrow[n\to\infty]{\text{a.s/}\mathbb{P}/(d)} f(T).$$

Continuous Mapping Theorem

Example: Recall that by LLN, $\bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\text{ a.s.}} p$. Therefore

$$f(\bar{R}_n) \xrightarrow[n\to\infty]{\mathbb{P},\text{ a.s.}} f(p) \quad \text{for any continuous } f$$

(Only need f to be continuous around p: f(x)=1/x works if $p > 0$)

We also have by CLT: $\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{(d)} Z$, $Z \sim \mathcal{N}(0,1)$. So

$$f\left(\sqrt{n}\,(\bar{R}_n - p)\right) \xrightarrow[n\to\infty]{(d)} f(Y) \quad Y \sim \mathcal{N}(0, p(1-p))$$

⚠ not the limit of $\sqrt{n}[f(\bar{R}_n) - f(p)]$!!

Delta (Δ) me

Recap

- Averages of random variables occur naturally in statistics
- We make modeling assumptions to apply probability results
- For large sample size they are *consistent* (LLN) and we know their distribution (CLT)
- CLT gives the (weakest) convergence in distribution but is enough to compute probabilities
- We use standardization and Gaussian tables to compute probabilities and quantiles
- We can make operations (addition, multiplication, continuous functions) on sequences of random variables

18.650 – Fundamentals of Statistics

2. Foundations of Inference

Goals

In this unit, we introduce a mathematical formalization of statistical modeling to make a principled sense of the **Trinity of statistical inference**.

We will make sense of the following statements:

1. Estimation:

   "$\hat{p} = \overline{R_n}$ is an estimator for the proportion p of couples that turn their head to the right"

   (side question: is 64.5% also an estimator for p?)

2. Confidence intervals:

   "$[0.56, 0.73]$ is a 95% confidence interval for p"

3. Hypothesis testing:

   "We find statistical evidence that more couples turn their head to the right when kissing"

The rationale behind statistical modeling

- Let $X_1, \dots, X_n$ be n independent copies of X.
- The goal of statistics is to learn the distribution of X.
- If $X \in \{0, 1\}$, easy! It's Bernoulli and we only have to learn the parameter p
- Can be more complicated. For example, here is a (partial) dataset with number of siblings (including self) that were collected from college students a few years back: 2, 3, 2, 4, 1, 3, 1, 1, 1, 1, 1, 2, 2, 3, 2, 2, 2, 3, 2, 1, 3, 1, 2, 3, ...
- We could make no assumption and try to learn the pmf:

x	1	2	3	4	5	6	≥ 7
$\mathbb{P}(X = x)$	p_1	p_2	p_3	p_4	p_5	p_6	$\sum_{i \geq 7} p_i$

  That's 7 parameters to learn. $\mathbb{P}(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}$

- Or we could assume that $X - 1 \sim \text{Poiss}(\lambda)$. That's 1 parameter to learn!

Statistical model

Formal definition

Let the observed outcome of a statistical experiment be a sample $X_1, \dots, X_n$ of n i.i.d. random variables in some measurable space E (usually $E \subseteq \mathbb{R}$) and denote by $\mathbb{P}$ their common distribution. A statistical model associated to that statistical experiment is a pair

$$(E, (\mathbb{P}_\theta)_{\theta \in \Theta}),$$

where:

- E is called sample space
- $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is a family of probility measures on E;
- Θ is any set, called parameter set

Parametric, nonparametric and semiparametric models

- Usually, we will assume that the statistical model is *well specified*, i.e., defined such that $\exists \theta$ such that $\mathbb{P} = \mathbb{P}_\theta$
- This particular θ is called the true parameter, and is unknown: The aim of the statistical experiment is to θ, or check it's properties when they have a special meaning $\theta > 2?$, $\theta \neq 1/2?$, $\dots$)
- We often assume that $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$: The model is called parametric
- Sometime we could have Θ be infinite dimensional in which case the model is called nonparametric
- If $\Theta = \Theta_1 \times \Theta_2$, where Θ_1 is finite dimensional and Θ_2 is infinite dimensional: *semiparametric* model. In these models we only care to estimate the finite dimensional parameter and the infinite dimensional one is called *nuisance* parameter. We will not cover such models in this class.

(θ, f)

Examples of parametric models

1. For n Bernoulli trials:
$$\left(\{0,1\}^n, (\mathrm{Ber}(p))_{p\in(0,1)}\right).$$
2. If $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Poiss}(\lambda)$ for some unknown $\lambda > 0$,
$$\left(\mathbb{N}^n, (\mathrm{Poiss}(\lambda))_{\lambda>0}\right).$$
3. If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$:
$$\left(\mathbb{R}^n, (\mathcal{N}(\mu, \sigma^2))_{(\mu,\sigma^2)\in \mathbb{R}\times(0,\infty)}\right).$$
4. If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}_d(\mu, I_d)$, for some unknown $\mu \in \mathbb{R}^d$:
$$\left(\mathbb{R}^{dn}, (\mathcal{N}_d(\mu, I_d))_{(\mu\in \mathbb{R}^d)}\right).$$

Examples of nonparametric models

1. If $X_1, \dots, X_n \in \mathbb{R}$ are i.i.d with unknown *unimodal*[1] pdf f:
$$E = \mathbb{R} \qquad \Theta = \{\text{unimodal pdf } f\}$$
$\mathbb{P}_\theta = \mathbb{P}_f$ = distribution with pdf f
2. If $X_1, \dots, X_n \in [0,1]$ are i.i.d with unknown invertible cdf F.
$E = [0,1]$ ...

[1] Increases on $(-\infty, a)$ and then decreases on (a, ∞) for some $a > 0$

Further examples

Sometimes we do not have simple notation to write $(\mathbb{P}_\theta)_{\theta\in\Theta}$, e.g., $(\mathrm{Ber}(p))_{p\in(0,1)}$ and we have to be more explicit:

1. **Linear regression model:** If $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d from the linear regression model $Y_i = \beta^\top X_i + \varepsilon_i$ $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$ for an unknown $\beta \in \mathbb{R}^d$ and $X_i \sim \mathcal{N}_d(0, I_d)$ independent of ε_i
$$E = \mathbb{R}^d \times \mathbb{R} \qquad \Theta = \mathbb{R}^d$$
2. **Cox proportional Hazard model:** If $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$: the conditional distribution of Y given $X = x$ has CDF F of the form
$$F(t) = 1 - \exp\left(-\int_0^t h(u) e^{(\beta^\top x)} du\right)$$
where h is an unknown non-negative nuisance function and $\beta \in \mathbb{R}^d$ is the parameter of interest.

Identifiability

The parameter θ is called *identifiable* iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.,

$$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently:

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

Examples

1. In all previous examples, the parameter is identifiable.
2. If $X_i = \mathbb{1}_{Y_i \geq 0}$ (indicator function), $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, are unobserved: μ and σ^2 are not identifiable (but $\theta = \mu/\sigma$ is).

Exercises

a) Which of the following is a statistical model?

1. $\left(\{1\}, (\text{Ber}(p))_{p \in (0,1)}\right)$
2. $\left(\{0,1\}, (\text{Ber}(p))_{p \in (0.2, 0.4)}\right)$ ✓
3. Both 1 and 2
4. None of the above

b) Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{U}([0,a])$ for some unknown $a > 0$. Which one of the following is the associated statistical model?

1. $\left([0,a], (\mathcal{U}([0,a]))_{a>0}\right)$
2. $\left(\mathbb{R}_+, (\mathcal{U}([0,a]))_{a>0}\right)$ ✓
3. $\left(\mathbb{R}, (\mathcal{U}([0,a]))_{a>0}\right)$
4. None of the above

uniform distribution on $[0,a]$

Exercises

c) Let $X_i = Y_i^2$, where $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{U}([0,a])$, for some unknown a, are unobserved. Is a identifiable?

1. Yes ✓
2. No

d) Let $X_i = \mathbb{1}_{Y_i \geq \frac{a}{2}}$, where $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{U}([0,a])$, for some unknown a, are unobserved. Is a identifiable?

1. Yes
2. No ✓

Estimation

Parameter estimation

Definitions

- *Statistic*: Any measurable[2] function of the sample, e.g., $\bar{X}_n, \max_i X_i$, $X_1 + \log(1 + |X_n|)$, sample variance, etc...
- *Estimator of* θ: Any statistic whose expression does not depend on θ
- An estimator $\hat{\theta}_n$ of θ is *weakly* (resp. *strongly*) consistent if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P} \text{ (resp. } a.s.)} \theta \quad \text{(w.r.t. } \mathbb{P}_\theta).$$

- An estimator $\hat{\theta}_n$ of θ is *asymptotically normal* if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$

The quantity σ^2 is then called *asymptotic variance* of $\hat{\theta}_n$.

[2] Rule of thumb: if you can compute it exactly once given data, it is measurable. You may have some issues with things that are implicitly defined

Bias of an estimator

- *Bias* of an estimator $\hat{\theta}_n$ of θ:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

- If $\text{bias}(\hat{\theta}) = 0$, we say that $\hat{\theta}$ is unbiased
- Example: Assume that $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$ and consider the following estimators for p:

  Bias $(\hat{p}_n) = \mathbb{E}[\hat{p}_n] - p$

  - $\hat{p}_n = \bar{X}_n$: $\text{bias}(\hat{p}_n) = 0$
  - $\hat{p}_n = X_1$: $\text{bias}(\hat{p}_n) = 0$
  - $\hat{p}_n = \dfrac{X_1 + X_2}{2}$: $\text{bias}(\hat{p}_n) = 0$
  - $\hat{p}_n = \sqrt{\mathbb{1}(X_1 = 1, X_2 = 2)}$ $\quad \hat{p}_n \sim \text{Ber}(p^2) \Rightarrow \text{bias}(\hat{p}_n) = p^2 - p$

    $Z \sim \text{Ber}(p^2)$

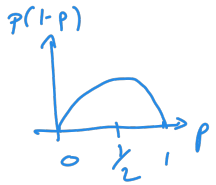
Variance of an estimator

$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

An estimator is a random variable so we can compute its variance.
In the previous examples:

$X \sim \text{Ber}(p)$
$\text{Var}(X) = p(1-p)$

- $\hat{p}_n = \bar{X}_n$: $\text{var}(\hat{p}_n) = \dfrac{p(1-p)}{n}$
- $\hat{p}_n = X_1$: $\text{var}(\hat{p}_n) = p(1-p)$
- $\hat{p}_n = \dfrac{X_1 + X_2}{2}$: $\text{Var}(\hat{p}_n) = \dfrac{p(1-p)}{2}$



- $\hat{p}_n = \sqrt{\mathbb{1}(X_1 = 1, X_2 = 2)}$ $\quad \text{Var}(\hat{p}_n) = p^2(1 - p^2)$

  $\hat{p}_n \sim \text{Ber}(p^2)$

Quadratic risk

- We want estimators to have low bias and low variance at the same time.
- The *Risk* (or *quadratic risk*) of an estimator $\hat{\theta}_n \in \mathbb{R}$ is

$$R(\hat{\theta}_n) = \mathbb{E}\left[|\hat{\theta}_n - \theta|^2\right]$$

Var $(\hat{\theta}_n)$ $\quad$ bias$(\hat{\theta}_n)^2$

$\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta)^2] = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta)^2]$
$+ 2\,\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta)]$

- Low quadratic risk means that both bias and variance are small:

  quadratic risk = VARIANCE + BIAS^2

Exercises

Let $X_1, X_2 \ldots, X_n$ be a random sample from $\mathcal{U}([a, a+1])$.
Questions a), b), c) and d) are about this sample.

a) Find $\mathbb{E}\left[\bar{X}_n\right] = a + \frac{1}{2}$

b) Is $\bar{X}_n - \frac{1}{2}$ an unbiased estimator for a? Yes : $\mathbb{E}\left[\bar{X}_n - \frac{1}{2}\right] = a$

c) Find the variance of $\bar{X}_n - \frac{1}{2}$. $\mathrm{Var}\left(\bar{X}_n - \frac{1}{2}\right) = \frac{1}{12n}$

d) Find the quadratic risk of $\bar{X}_n - \frac{1}{2}$.

$$R\left(\bar{X}_n - \frac{1}{2}\right) = \frac{1}{12n} + 0^2 = \frac{1}{12n}$$

Confidence intervals

Confidence intervals

Let $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model based on observations $X_1, \ldots, X_n$, and assume $\Theta \subseteq \mathbb{R}$. Let $\alpha \in (0, 1)$.

- *Confidence interval (C.I.) of level* $1-\alpha$ for θ: Any random (depending on $X_1, \ldots, X_n$) interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:

$$\mathbb{P}_\theta\left[\mathcal{I} \ni {}^3\theta\right] \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

- *C.I. of asymptotic level* $1-\alpha$ for θ: Any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:

$$\lim_{n\to\infty} \mathbb{P}_\theta\left[\mathcal{I} \ni \theta\right] \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

[3] $\mathcal{I} \ni \theta$ means that $\mathcal{I}$ contains θ. This notation emphasizes the randomness of $\mathcal{I}$ but we can equivalently write $\theta \in \mathcal{I}$.

A confidence interval for the kiss example

- Recall that we observe $R_1, \ldots, R_n \overset{iid}{\sim} \mathrm{Ber}(p)$ for some unknown $p \in (0, 1)$.
- Statistical model: $\left(\{0,1\}, (\mathrm{Ber}(p))_{p\in(0,1)}\right)$. ✓
- Recall that our estimator for p is $\hat{p} = \bar{R}_n$. ✓
- From CLT:

$$\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

This means (precisely) that:

- $\Phi(x)$: cdf of $\mathcal{N}(0,1)$; $\Phi_n(x)$: cdf of $\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{p(1-p)}}$.
- Then: $\Phi_n(x) \approx \Phi(x)$ (CLT) when n becomes large. Hence, for all $x > 0$,

$$\mathbb{P}\left[|\bar{R}_n - p| \geq x\right] \simeq 2\left(1 - \Phi\left(\frac{x\sqrt{n}}{\sqrt{p(1-p)}}\right)\right).$$

Confidence interval?

- For a fixed $\alpha \in (0,1)$, if $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$, then with probability $\simeq 1-\alpha$ (if n is large enough !),

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, p + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \right].$$

- It yields

$$\lim_{n\to\infty} \mathbb{P}\left(\left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \right] \ni p \right) = 1-\alpha$$

- But this is **not** a confidence interval because it depends on p!

- To fix this, there are 3 solutions.

Solution 1: Conservative bound

- Note that no matter the (unknown) value of p,

$$p(1-p) \leq \frac{1}{4}$$

- Hence, roughly with probability at least $1-\alpha$,

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}}{2\sqrt{n}}, p + \frac{q_{\alpha/2}}{2\sqrt{n}} \right].$$

- We get the asymptotic confidence interval:

$$\mathcal{I}_{\text{conserv}} = \left[\bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}} \right]$$

- Indeed

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{I}_{\text{conserv}} \ni p) \geq 1-\alpha$$

Solution 2: Solving the (quadratic) equation for p

- We have the system of two inequalities in p:

$$\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

- Each is a quadratic inequality in p of the form

$$(p-\bar{R}_n)^2 \leq \frac{q_{\alpha/2}^2 p(1-p)}{n}$$

We need to find the roots $p_1 < p_2$ of

$$(1+\frac{q_{\alpha/2}^2}{n})p^2 - (2\bar{R}_n + \frac{q_{\alpha/2}^2}{n})p + \bar{R}_n^2 = 0$$

- This leads to a new confidence interval $\mathcal{I}_{\text{solve}} = [p_1, p_2]$ such that:

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{I}_{\text{solve}} \ni p) = 1-\alpha$$

(it's complicated to write in generic way so let us wait to have values for n, α and $\bar{R}_n$ to plug-in)

Solution 3: plug-in

- Recall that by the LLN $\hat{p} = \bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\text{a.s.}} p$
- So by Slutsky, we also have

$$\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

- This leads to a new confidence interval:

$$\mathcal{I}_{\text{plug-in}} = \left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right]$$

such that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{I}_{\text{plug-in}} \ni p) = 1-\alpha$$

95% asymptotic CI for the kiss example

Recall that in the kiss example we had $n = 124$ and $\bar{R}_n = 0.645$.
Assume $\alpha = 5\%$.
For $\mathcal{I}_{\text{solve}}$, we have to find the roots of:

$$1.03p^2 - 1.32p + 0.41 = 0 \qquad p_1 = 0.53, p_2 = 0.75$$

We get the following confidence intervals of asymptotic level 95%:

- $\mathcal{I}_{\text{conserv}} = [0.56\,,\, 0.73]$
- $\mathcal{I}_{\text{solve}} = [0.53\,,\, 0.75]$
- $\mathcal{I}_{\text{plug-in}} = [0.56\,,\, 0.73]$

There are many[4] other possibilities in softwares even ones that use the exact distribution of $n\bar{R}_n \sim$ Bin(n, p)

$$\mathcal{I}_{\text{R default}} = [0.55\,,\, 0.73]$$

[4] See R. Newcombe (1998). *Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods*

Exercises

a) Let I, J be some 95% and 98% asymptotic confidence intervals (respectively) for p. Which one of the following statements is correct?

1. We always have $I \subset J$.
2. We always have $J \subset I$.
3. None of the above.

b) Find a 98% asymptotic confidence interval for p.

1−α C.I. $\quad \bar{R}_n \pm \frac{q_{\alpha/2}}{2\sqrt{n}}$ $\quad q_{1\%} = 2.33$(?) (Table check!)

$\left[0.645 \pm \frac{2.33}{2\sqrt{124}}\right]$

Exercises

c) Consider a new experiment in which there are 150 participants, 75 turned left and 75 turned right. Which of the following is the correct answer?

1. $[0, 0.5]$ is a 50% asymptotic confidence intervals for p
2. $[0.5, 1]$ is a 50% asymptotic confidence intervals for p
3. $[0.466, 0.533]$ is a 50% asymptotic confidence intervals for p
4. $[0.48, 0.52]$ is a 50% asymptotic confidence intervals for p
5. both (1) and (2)
6. (1), (2) and (3)
7. (1), (2), (3) and (4)

$\mathbb{P}\left[\frac{\sqrt{n}(\bar{R}_n - p)}{\sqrt{p(1-p)}} \leq 0\right] \xrightarrow[n\to\infty]{} \frac{1}{2} = 50\%$

$\bar{R}_n = 0.5$ $\quad \mathbb{P}[p \geq \bar{R}_n] \xrightarrow[n\to\infty]{} 50\%$

$[\bar{R}_n, 1]$ is a 50% asymptotic CI.

$[0, \bar{R}_n]$ also because $\mathbb{P}(\bar{R}_n - p \geq 0) \xrightarrow[n\to\infty]{} \frac{1}{2}$

Exercises

d) If $[0.34, 0.57]$ is a 95% confidence interval for an unknown proportion p, then the probability that p is in this interval is at least asymptotically

1. 0.025
2. 0.05
3. 0.95
4. None of the above

e) If $[0.34, 0.57]$ is a 95% confidence interval for an unknown proportion p, is it also a 98% confidence interval?

1. Yes
2. No

f) If $[0.34, 0.57]$ is a 95% confidence interval for an unknown proportion p, is it also a 90% confidence interval?

1. Yes
2. No

Another example: The T

Slide # 29
Object Source / URL*
https://www.youtube.com/watch?v=VBBeRDa_gms
Citation/Attribution -- Photo from User: Iconic Rails on YouTube (c) Mikay Royce

Statistical problem

- You observe the times (in minutes) between arrivals of the T at Kendall: $T_1, \ldots, T_n$.
- You **assume** that these times are:
  - Mutually independent ✓
  - Exponential random variables with common parameter ✓
- You want to *estimate* the value of λ, based on the observed arrival times.

Discussion of the modeling assumptions

- Mutual independence of $T_1, \ldots, T_n$: plausible but not completely justified (often the case with independence).
- $T_1, \ldots, T_n$ are exponential r.v.: **lack of memory** of the exponential distribution:

$$\mathbb{P}[T_1 > t + s | T_1 > t] = \mathbb{P}[\qquad], \quad \forall s, t \geq 0.$$

  Also, $T_i > 0$ almost surely!
- The exponential distributions of $T_1, \ldots, T_n$ have the same parameter: in average all the same inter-arrival time. True only for limited period (rush hour $\neq$ 11pm).

Estimator

- Density of T_1:

$$f(t) = \lambda e^{-\lambda t}, \quad \forall t \geq 0.$$

- $\mathbb{E}[T_1] = \frac{1}{\lambda}$.
- Hence, a natural estimate of $\frac{1}{\lambda}$ is

$$\bar{T}_n := \frac{1}{n} \sum_{i=1}^{n} T_i.$$

- A natural estimator of λ is

$$\hat{\lambda} := \frac{1}{\bar{T}_n} \xrightarrow[n \to \infty]{a.s., \mathbb{P}} \lambda$$

$\mathbb{E}\left[\frac{1}{T_n}\right] > \frac{1}{\mathbb{E}[T_n]} = \lambda$

$\lambda > 0$

First properties

- By the LLN's,
$$\bar{T}_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \frac{1}{\lambda}$$
- Hence,
$$\hat{\lambda} \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \lambda.$$
- By the CLT,
$$\sqrt{n}\left(\bar{T}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{} \mathcal{N}(0, \lambda^{-2}).$$
- How does the CLT transfer to $\hat{\lambda}$? How to find an asymptotic confidence interval for λ ?

The Delta method

Let $(Z_n)_{n\geq 1}$ sequence of r.v. that satisfies
$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$
for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ (the sequence $(Z_n)_{n\geq 1}$ is said to be *asymptotically normal around* θ).

Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at the point θ. Then,
- $(g(Z_n))_{n\geq 1}$ is also asymptotically normal;
- More precisely,
$$\sqrt{n}\,(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

Consequence of the Delta method

- $\sqrt{n}\left(\hat{\lambda} - \lambda\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \lambda^2)$.
- Hence, for $\alpha \in (0,1)$ and when n is large enough,
$$|\hat{\lambda} - \lambda| \leq \lambda \cdot \frac{q_{\alpha/2}}{\sqrt{n}}$$
with probability approximately $1-\alpha$.
- Can $\left[\hat{\lambda} - \frac{q_{\alpha/2}\lambda}{\sqrt{n}}, \hat{\lambda} + \frac{q_{\alpha/2}\lambda}{\sqrt{n}}\right]$ be used as an asymptotic confidence interval for λ ?

No : depends on λ

Three solutions

1. The conservative bound: we have no a priori way to bound λ
2. We can solve for λ:
$$|\hat{\lambda} - \lambda| \leq \frac{q_{\alpha/2}\lambda}{\sqrt{n}} \iff \lambda\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right) \leq \hat{\lambda} \leq \lambda\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)$$
$$\iff \frac{\hat{\lambda}}{1 + \frac{q_{\alpha/2}}{\sqrt{n}}} \leq \lambda \leq \frac{\hat{\lambda}}{1 - \frac{q_{\alpha/2}}{\sqrt{n}}}$$
It yields
$$\mathcal{I}_{\text{solve}} = \left[\hat{\lambda}\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}, \hat{\lambda}\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}\right]$$
3. Plug-in yields
$$\mathcal{I}_{\text{plug-in}} = \left[\hat{\lambda}\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right), \hat{\lambda}\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)\right]$$

95% asymptotic CI for the T example

Assume that $n = 64$ and $\bar{T}_n = 6.23$ and $\alpha = 5\%$.

We get the following confidence intervals of asymptotic level 95%:

- $\mathcal{I}_{\text{solve}} = [0.13\,,\,0.21]$
- $\mathcal{I}_{\text{plug-in}} = [0.12\,,\,0.20]$

Hypothesis testing

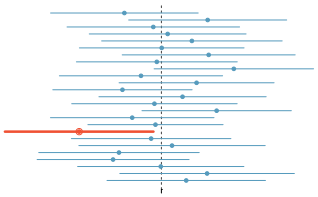
Meaning of a confidence interval

Take $\mathcal{I}_{\text{plug-in}} = [0.12\,,\,0.20]$ for example. What is the meaning of "$\mathcal{I}_{\text{plug-in}}$ is a confidence intervals of asymptotic level 95%".

Does it mean that

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda \in \quad [0.12\,,\,0.20]\right) \geq .95?$$ No

There is a *frequentist* interpretation[5]: If we were to repeat this experiment (collect 64 observations) then λ would be in the resulting confidence interval about 95% of the time (image credit: openintro.org).



[5] The frequentist approach is often contrasted with the Bayesian approach

How to board a plane?

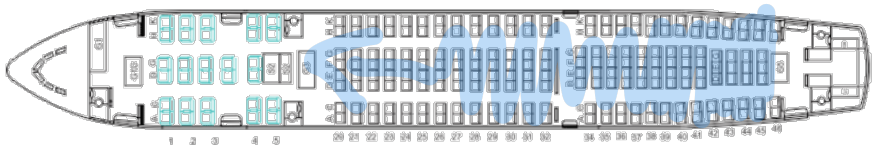
ect Source / URL*
https://www.ewg.org/release/welcome-aboard-coach-class-scott-pruitt-here-s-your-ewg-approved-travel-kit
Citation/Attribution -- Image Copyright © 2019, Environmental Working Group. All rights reserved.

What is the fastest boarding method?

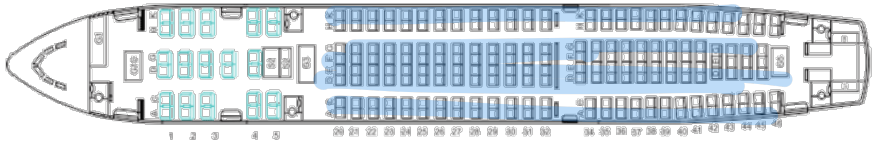
What is the fastest method to board a plane?

R2F or WilMA?

- R2F= Rear to Front



- WilMA=Window, Middle, Aisle. It is basically an OUTSIDE to INSIDE method.



The data

We collected data from two different airlines: JetBlue (R2F) and United (WilMA).
We got the following results:

	R2F	WilMA
Average (mins)	24.2	15.9
Std. Dev (mins)	2.1	1.3
Sample size	72	56

Model and Assumptions

- Let X (resp. Y) denote the boarding time of a random JetBlue (resp. United) flight.
- We *assume* that $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- Let n and m denote the JetBlue and United sample sizes respectively.
- We have $X_1, \ldots, X_n$ independent copies of X and $Y_1, \ldots Y_m$ independent copies of Y.
- We further assume that the two samples are independent.

We want to answer the question:

$$\text{Is } \mu_1 = \mu_2 \text{ or is } \mu_1 > \mu_2?$$

By making **modeling assumptions**, we have reduced the number of ways the hypothesis $\mu_1 = \mu_2$ may be rejected. We do not allow that $\mu_1 < \mu_2$!

We have two samples: this is a **two-sample test**

A first heuristic

Simple heuristic:

$$\text{"If } \bar{X}_n > \bar{Y}_m \text{, then } \mu_1 > \mu_2\text{"}$$

This could go wrong if I randomly pick only full flights in my sample $X_1, \ldots, X_n$ and empty flights in my sample $Y_1, \ldots, Y_m$.

Better heuristic:

$$\text{"If } \bar{X}_n - \text{Buffer}_n > \bar{Y}_m + \text{Buffer}_m \text{ then } \mu_1 > \mu_2\text{"}$$

To make this intuition more precise, we need to take the size of the random fluctuations of $\bar{X}_n$ and $\bar{Y}$ into account!

Waiting time in the ER

- The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- Here, we collect only one sample: $X_1, \dots, X_n$ (waiting time in minutes for n random patients) with unknown expected value $\mathbb{E}[X_1] = \mu$.
- We want to know if $\mu > 30$.

This is a **one-sample test**

Object Source / URL*
https://doctors.healthtap.com/hs-fs/hubfs/waiting_room.jpg?t=1511990105503&width=640&name=waiting_room.jpg
Citation/Attribution -- Image © 2018 HealthTap, All rights reserved

Heuristic

Heuristic:

"If

$$\bar{X}_n + \text{Buffer}_n < 30$$

then conclude that $\mu \leq 30$ "

Example 1

According to a survey conducted in 2017 on 4,971 randomly sampled Americans, 32% report to get at least some of their news on Youtube. Can we conclude that at most a third of all Americans get at least some of their news on Youtube?

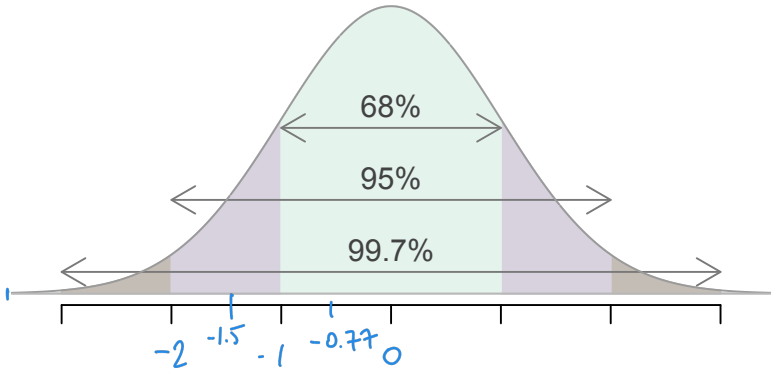
- $n = 4,971$, $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$;
- $\bar{X}_n = 0.32$
- If it was true that $p = .33$: By CLT,

$$\mathbb{E}[\bar{X}_n] = .33 \qquad \text{Var}[\bar{X}_n] = \frac{.33(1-.33)}{4,971}$$

$$\sqrt{n}\frac{\bar{X}_n - .33}{\sqrt{.33(1-.33)}} \approx \mathcal{N}(0,1).$$

- $\sqrt{n}\frac{\bar{X}_n - .33}{\sqrt{.33(1-.33)}} \approx -1.50$
- Conclusion:

The Standard Gaussian distribution



Example 2

Example 2: A coin is tossed 30 times, and Heads are obtained 13 times. Can we conclude that the coin is significantly unfair ?

- $n = 30, X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$;
- $\bar{X}_n = 13/30 \approx .43$
- If it was true that $p = .5$: By CLT,

$$\sqrt{n}\frac{\bar{X}_n - .5}{\sqrt{.5(1-.5)}} \approx \mathcal{N}(0,1).$$

- Our data gives $\sqrt{n}\dfrac{\bar{X}_n - .5}{\sqrt{.5(1-.5)}} \approx -0.77$
- The number $-.77$ is a plausible realization of a random variable $Z \sim \mathcal{N}(0,1)$.
- Conclusion: It is unlikely that the coin is unfair

Asymmetry in the hypotheses

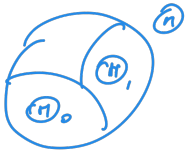
- H_0 and H_1 do not play a symmetric role: the data is is only used to try to disprove H_0
- In particular lack of evidence, does not mean that H_0 is true ("innocent until proven guilty")
- A *test* is a statistic $\psi \in \{0,1\}$ such that:
  - If $\psi = 0$, H_0 is not rejected;
  - If $\psi = 1$, H_0 is rejected. $\Leftrightarrow H_1$

$\psi(x) = \mathbb{1}(\psi(x) = 1)$

- Coin example: H_0: $p = 1/2$ vs. H_1: $p \neq 1/2$.
- $\psi = \mathbb{1}\left\{ \sqrt{n}\dfrac{|\bar{X}_n - \frac{1}{2}|}{\sqrt{0.5(1-0.5)}} > C \right\}$, for some $C > 0$.
- How to choose the *threshold* C ?

Statistical formulation

- Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Let Θ_0 and Θ_1 be disjoint subsets of Θ.



- Consider the two hypotheses: $\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$
- H_0 is the *null hypothesis*, H_1 is the *alternative hypothesis*.
- If we believe that the true θ is either in Θ_0 or in Θ_1, we may want to *test H_0 against H_1*.
- We want to decide whether to *reject H_0* (look for evidence against H_0 in the data).

Errors

- *Rejection region* of a test ψ: where $(X_1, \dots, X_n)$ lives

$$R_\psi = \{x \in E^n : \psi(x) = 1\}. \qquad \psi(x) = \mathbb{1}(x \in R_\psi)$$

- *Type 1 error* of a test ψ (rejecting H_0 when it is actually true):

$$\begin{array}{rccc} \alpha_\psi & : & \Theta_0 & \to & \mathbb{R} \\ & & \theta & \mapsto & \mathbb{P}_\theta[\psi = 1]. \end{array} \qquad (\text{or } [0,1])$$

- *Type 2 error* of a test ψ (not rejecting H_0 although H_1 is actually true):

$$\begin{array}{rccc} \beta_\psi & : & \Theta_1 & \to & \mathbb{R} \\ & & \theta & \mapsto & \mathbb{P}_\theta[\psi = 0] \end{array}$$

- *Power* of a test ψ:

$$\pi_\psi = \inf_{\theta \in \Theta_1} (1 - \beta_\psi(\theta)).$$

Level, test statistic and rejection region

- A test ψ has *level* α if (I think $\alpha = 5\%, 1\%, \ldots$)

$$\alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- A test ψ has *asymptotic level* α if

$$\lim_{n\to\infty} \alpha_{\psi_n}(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- In general, a test has the form

$$\psi = \mathbb{1}\{T_n > c\},$$

($\psi = \mathbb{1}\{|T_n| > c\}$, $\psi = \mathbb{1}\{|T_n| \leq c\}$)

for some statistic T_n and threshold $c \in \mathbb{R}$.

- T_n is called the *test statistic*. The rejection region is $R_\psi = \{T_n > c\}$

One-sided vs two-sided tests

We can refine the terminology when $\theta \in \Theta \subset \mathbb{R}$ and H_0 is of the form

$$H_0 : \theta = \theta_0 \quad \Leftrightarrow \quad \Theta_0 = \{\theta_0\}$$

- If $H_1 : \theta \neq \theta_0$: **two-sided test** ✓
- If $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$: **one-sided test** ✓

Examples:

- Boarding method: one sided
- Waiting time in the ER: one sided
- The kiss example: two sided
- Fair coin: two sided

One or two sided tests will have different rejection regions.

Bernoulli experiment

- Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p)$, for some unknown $p \in (0, 1)$.
- We want to test:

$$H_0\text{: } p = 1/2 \text{ vs. } H_1\text{: } p \neq 1/2$$

with asymptotic level $\alpha \in (0, 1)$.

- Let $T_n = \left| \sqrt{n} \, \dfrac{\hat{p}_n - 0.5}{\sqrt{.5(1-.5)}} \right|$, where $\hat{p}_n$ is the MLE. ($\bar{X}_n$)
- If H_0 is true, then by CLT,

$$\mathbb{P}[T_n > q_{\alpha/2}] \xrightarrow[n\to\infty]{} 0.05$$

- Let $\psi_\alpha = \mathbb{1}\{T_n > q_{\alpha/2}\}$.

Examples

For $\alpha = 5\%$, $q_{\alpha/2} = 1.96$, ($q_\alpha = 1.645$)

Fair coin

H_0 is not rejected at the asymptotic level 5% by the test $\psi_{5\%}$. ($0.77 < 1.96$)

News on Youtube

$H_0 : p \geq 0.33$ vs. $H_1 : p < 0.33$. This is a one-sided test.

We reject if:

$$\sqrt{n} \, \frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < c$$

But what value for $p \in \Theta_0 = [0.33, 1]$ should we choose?

Type 1 error is the function $p \mapsto \mathbb{P}_p[\psi = 1]$. To control the level we need to find the p that **maximizes** it over Θ_0
$\rightarrow$ no need for computations, it's clearly $p = 0.33$ ($-1.645 < -1.50$)

H_0 is not rejected at the asymptotic level 5% by the test $\psi_{5\%}$.

p-value

Definition

The (asymptotic) *p-value* of a test ψ_α is the smallest (asymptotic) level α at which ψ_α rejects H_0. It is random, it depends on the sample.

Golden rule

p-value $\leq \alpha \quad \Leftrightarrow \quad H_0$ is rejected by ψ_α, at the (asymptotic) level α.

The smaller the p-value, the more confidently one can reject H_0.

Exercise: kiss

Recall that in the Kiss example we observed 80 out of 124 couples turning their head to the right. Formulate the statistical hypothesis problem, compute the p-value and conclude.

Exercise: Cookies[6]

Students are asked to count the number of chocolate chips in 32 cookies for a class activity. They found that the cookies on average had 14.77 chocolate chips with a standard deviation of 4.37 chocolate chips. The packaging for these cookies claims that there are at least 20 chocolate chips per cookie. One student thinks this number is unreasonably high since the average they found is much lower. Another student claims the difference might be due to chance. What do you think (compute a p-value)?

Object Source / URL* https://i-love-png.com/tags/cookie-dough.html Citation/ Attribution -- Image from Free PNG Library © 2019

[6] from the textbook OpenIntro Statistics

Exercise : Machine learning predicts breast cancer

A vast problem in breast cancer are false positive, that is surgery performed on benign tumors. A new machine learning procedure claims to improve the state-of-the art (95% of false positive) significantly while preserving the same true positive rate (detecting malignant tumors as malignant). To verify this claim, we collected data on 297 benign tumors. The algorithm recommended to perform surgery on 206 of them.

Let p denote the proportion of benign tumors on which the algorithm prescribes surgery.
Formulate the statistical hypothesis problem, compute the p-value and conclude.

Recap

- A statistical model is a pair of the form $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ where E is the sample space and $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a family of candidate probability distributions.
- A model can be well specified and identifiable.
- The trinity of statistical inference: estimation, confidence intervals and testing
- Estimator: one value whose performance can be measured by consistency, asymptotic normality, bias, variance and quadratic risk
- Confidence intervals provide "error bars" around estimators. Their size depends on the confidence level
- Hypothesis testing: we want to ask a yes/no answer about an unknown parameter. They are characterized by hypotheses, level, power, test statistic and rejection region. Under the null hypothesis, the value of the unknown parameter becomes known (no need for plug-in).

Unit 4
Slide # 9
Object Source / URL*
https://thesciencedog.com/2017/01/12/consider-the-placebo-sugar-pills-and-your-dog/
Citation/Attribution -- Cartoon by Chris Madden © CartoonStock Ltd. 2019All Rights Reserved.

Unit 4
Slide # 22
Object Source / URL*
https://medium.com/value-stream-design/the-curious-tale-of-william-sealy-gosset-b3178a9f6ac8
Citation/Attribution - Image from article on Medium (c) Max Pagels.

Unit 4
Slide # 23
Object Source / URL*
https://medium.com/value-stream-design/the-curious-tale-of-william-sealy-gosset-b3178a9f6ac8
Citation/Attribution-- Image (c) Medium Corporation; all rights reserved.

Unit 4
Slide #23
Object Source / URL*
https://www.irishexaminer.com/breakingnews/lifestyle/international-stout-day-7-things-you-need-to-know-about-this-dark-beer-882510.html
Citation/Attribution-- Image © Irish Examiner Ltd

Unit 4
Slide # 59c.
Object Source / URL*
https://cruel.org/econthought/profiles/wald.html;
https://commons.wikimedia.org/wiki/File:Abraham_Wald.jpg
Citation/Attribution -- Image on Wikimedia by Konrad Jacobs, Erlangen, Copyright is MFO. (CC BY-SA) 2.0

Unit 4
Slide # 59b.
Object Source / URL*
http://webdept.fiu.edu/~statdept/ssss.htm
Citation/Attribution -- Photo (c) Samuel Shapiro.

Unit 4
Slide # 59c.
Object Source / URL*
https://www.sapaviva.com/andrey-kolmogorov/
Citation/Attribution - Image from Sapaviva.com © 2000. 2017 Valentine O. Oduenyi

Unit 4
Slide # 59d.
Object Source / URL*
https://www.causeweb.org/cause/resources/fun/quotes/wellswilks-statistical-thinking
Citation/Attribution -- Image © 2019 Consortium for the Advancement of Undergraduate Statistics Education. (CC BY NC SA) 4.0

Unit 4
Slide # 59e.
Object Source / URL*
https://www.statmania.info/2017/09/harald-cramer.html Citation/Attribution -- Image (c) 2016 Stat Mania

Unit 4
Slide # 59a.
Object Source / URL*
https://rai.ucar.edu/projects/extremes/Extremes/history.html;
https://photos.aip.org/history-programs/niels-bohr-library/photos/von-mises-richard-a12 Citation/Attribution -- Image from the AIP Emilio Segrè Visual Archives, Von Mises Collection. (c) © 2019 American Institute of Physics

18.650 – Fundamentals of Statistics

3. Methods for estimation

Distance measures probability distributions

Goals

In the kiss example, the estimator was **intuitively** the right thing to do: $\hat{p} = \bar{X}_n$.

In view of LLN, since $p = \mathbb{E}[X]$, we have $\bar{X}_n$ so $\hat{p} \approx p$ for n large enough.

1. Maximum likelihood estimation (MLE): a generic approach with very good properties
2. Method of moments: a (fairly) generic and easy approach that extends the setup where $\theta = \mathbb{E}[X]$
3. M-estimators: a generalization of MLE, flexible, and close to machine learning

Total variation distance

Let $\left(E, (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \mathbb{P}_{\theta^*}$: θ^* is the **true** parameter.

Statistician's goal: given $X_1, \dots, X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$ for the true parameter θ^*.
This means: $\left|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A)\right|$ is **small** for all $A \subset E$.

Definition

The *total variation distance* between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A \subset E} \left|\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A) \right|.$$

Total variation distance between discrete measures

Assume that E is discrete (i.e., finite or countable). This includes Bernoulli, Binomial, Poisson, ...

Therefore X has a PMF (probability mass function): $\mathbb{P}_\theta(X = x) = p_\theta(x)$ for all $x \in E$,

$$p_\theta(x) \geq 0\,, \quad \sum_{x \in E} p_\theta(x) = 1\,.$$

The total variation distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is a simple function of the PMF's p_θ and $p_{\theta'}$:

$$\mathsf{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{x \in E} \big|p_\theta(x) - p_{\theta'}(x)\big|\,.$$

Total variation distance between continuous measures

Assume that E is continuous. This includes Gaussian, Exponential, ...

Assume that X has a density $\mathbb{P}_\theta(X \in A) = \int_A f_\theta(x)dx$ for all $A \subset E$.

$$f_\theta(x) \geq 0, \quad \int_E f_\theta(x)dx = 1\,.$$

The total variation distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is a simple function of the densities f_θ and $f_{\theta'}$:

$$\mathsf{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int_E \big|f_\theta(x) - f_{\theta'}(x)\big|dx\,.$$

An estimation strategy

Build an estimator $\widehat{\mathsf{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ for all $\theta \in \Theta$. Then find $\hat{\theta}$ that *minimizes* the function $\theta \mapsto \widehat{\mathsf{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.

problem: Unclear how to build $\widehat{\mathsf{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$!

Kullback-Leibler (KL) divergence

There are **many** distances between probability measures to replace total variation. Let us choose one that is more convenient.

Definition

The *Kullback-Leibler*[1] *(KL) divergence* between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

$$\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \displaystyle\sum_{x \in E} p_\theta(x) \log\Big(\frac{p_\theta(x)}{p_{\theta'}(x)}\Big) & \text{if } E \text{ is discrete} \\ \displaystyle\int_E f_\theta(x) \log\Big(\frac{f_\theta(x)}{f_{\theta'}(x)}\Big)dx & \text{if } E \text{ is continuous} \end{cases}$$

[1] KL divergence is also know as "relative entropy"

Properties of KL-divergence

- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \mathsf{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- If $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \not\leq \mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \mathsf{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Not a distance.

This is is called a *divergence*.

Asymmetry is the key to our ability to estimate it!

Maximum likelihood estimation

Estimating the KL

$$\mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*}\Big[\log\Big(\frac{p_{\theta^*}(X)}{p_\theta(X)}\Big)\Big]$$

$$= \mathbb{E}_{\theta^*}\big[\log p_{\theta^*}(X)\big] - \mathbb{E}_{\theta^*}\big[\log p_\theta(X)\big]$$

So the function $\theta \mapsto \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is of the form:

$$\text{“constant”} - \mathbb{E}_{\theta^*}\big[\log p_\theta(X)\big]$$

Can be estimated: $\mathbb{E}_{\theta^*}[h(X)] \rightsquigarrow \frac{1}{n}\sum_{i=1}^n h(X_i)$ (by LLN)

$$\widehat{\mathsf{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{“constant”} - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

Maximum likelihood

$$\widehat{\mathsf{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{“constant”} - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

$$\begin{aligned}
\min_{\theta\in\Theta} \widehat{\mathsf{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) &\quad\Leftrightarrow\quad \min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \\
&\quad\Leftrightarrow\quad \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \\
&\quad\Leftrightarrow\quad \max_{\theta\in\Theta} \sum_{i=1}^n \log p_\theta(X_i) \\
&\quad\Leftrightarrow\quad \max_{\theta\in\Theta} \prod_{i=1}^n p_\theta(X_i)
\end{aligned}$$

This is the **maximum likelihood principle**.

Likelihood, Discrete case (1)

Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that E is discrete (i.e., finite or countable).

Definition

The *likelihood* of the model is the map L_n (or just L) defined as:

$$\begin{array}{rccc} L_n & : & E^n \times \Theta & \to & \mathbb{R} \\ & & (x_1, \dots, x_n; \theta) & \mapsto & \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]. \end{array}$$

Likelihood for the Bernoulli model

Example 1 (Bernoulli trials): If $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$ for some $p \in (0,1)$:

- $E = \{0, 1\}$;
- $\Theta = (0, 1)$;
- $\forall (x_1, \dots, x_n) \in \{0,1\}^n, \quad \forall p \in (0,1),$

$$\begin{aligned} L(x_1, \dots, x_n; p) &= \prod_{i=1}^n \mathbb{P}_p[X_i = x_i] \\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

Likelihood for the Poisson model

Example 2 (Poisson model):
If $X_1, \dots, X_n \overset{iid}{\sim} \text{Poiss}(\lambda)$ for some $\lambda > 0$:

- $E = \mathbb{N}$;
- $\Theta = (0, \infty)$;
- $\forall (x_1, \dots, x_n) \in \mathbb{N}^n, \quad \forall \lambda > 0,$

$$L(x_1, \dots, x_n; \lambda) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \dots x_n!}.$$

Likelihood, Continuous case

Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that all the $\mathbb{P}_\theta$ have density f_θ.

Definition

The *likelihood* of the model is the map L defined as:

$$\begin{array}{rccc} L & : & E^n \times \Theta & \to & \mathbb{R} \\ & & (x_1, \dots, x_n; \theta) & \mapsto & \prod_{i=1}^n f_\theta(x_i). \end{array}$$

Likelihood for the Gaussian model

Example (Gaussian model): If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some $\mu \in \mathbb{R}, \sigma^2 > 0$:

- $E = \mathbb{R}$;
- $\Theta = \mathbb{R} \times (0, \infty)$
- $\forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$,

$$L(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Likelihood for the Uniform model

Example (Uniform model): If $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, for some $\theta > 0$:

- $E = (0, \infty)$;
- $\Theta = (0, \infty)$
- $\forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$,

$$L(x_1, \dots, x_n; \theta) = \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} \leq \theta\}$$

where $x_{(n)} = \max_i x_i$

Likelihood for the Mixture of two Gaussians model

Example 1 (Mixture of Gaussians model): If $X_1, \dots, X_n$ are i.i.d from a mixture of two Gaussians, with means $\mu_1, \mu_2 \in \mathbb{R}$, variances, $\sigma_1^2, \sigma_2^2 > 0$ and $\pi \in (0, 1)$

- $E = \mathbb{R}$;
- $\Theta = \mathbb{R} \times \mathbb{R} \times (0, 1)$
- $\forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$,

$$L(x_1, \dots, x_n; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) =$$
$$\frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^{n} \left\{ \frac{\pi}{\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) + \frac{1-\pi}{\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right) \right\}.$$

Maximum likelihood estimator

Let $X_1, \dots, X_n$ be an i.i.d. sample associated with a statistical model $\left(E, (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$ and let L be the corresponding likelihood.

Definition

The *maximum likelihood estimator* of θ is defined as:

$$\hat{\theta}_n^{\text{MLE}} = \underset{\theta \in \Theta}{\text{argmax}} \; L(X_1, \dots, X_n, \theta),$$

provided it exists.

Remark (log-likelihood estimator): In practice, we use the fact that

$$\hat{\theta}_n^{\text{MLE}} = \underset{\theta \in \Theta}{\text{argmax}} \; \log L(X_1, \dots, X_n, \theta).$$

Interlude: maximizing/minimizing functions

Note that

$$\min_{\theta\in\Theta} -h(\theta) \quad \Leftrightarrow \quad \max_{\theta\in\Theta} h(\theta)$$

In this class, we focus on **maximization**.

Maximization of arbitrary functions can be difficult:

Example: $\theta \mapsto \prod_{i=1}^{n}(\theta - X_i)$

Concave and convex functions

Definition

A function twice differentiable function $h : \Theta \subset \mathbb{R} \to \mathbb{R}$ is said to be *concave* if its second derivative satisfies

$$h''(\theta) \leq 0\,, \qquad \forall\ \theta \in \Theta$$

It is said to be *strictly concave* if the inequality is strict: $h''(\theta) < 0$

Moreover, h is said to be (strictly) *convex* if $-h$ is (strictly) concave, i.e. $h''(\theta) \geq 0$ ($h''(\theta) > 0$).

Examples:

- $\Theta = \mathbb{R}$, $h(\theta) = -\theta^2$,
- $\Theta = (0, \infty)$, $h(\theta) = \sqrt{\theta}$,
- $\Theta = (0, \infty)$, $h(\theta) = \log\theta$,
- $\Theta = [0, \pi]$, $h(\theta) = \sin(\theta)$
- $\Theta = \mathbb{R}$, $h(\theta) = 2\theta - 3$

Optimality conditions

Strictly concave functions are easy to maximize: if they have a maximum, then it is **unique**. It is the unique solution to

$$h'(\theta) = 0\,,$$

or, in the multivariate case

$$\nabla h(\theta) = 0 \in \mathbb{R}^d\,.$$

There are many algorithms to find it numerically: this is the theory of "convex optimization". In this class, often a **closed form formula** for the maximum.

Examples of maximum likelihood estimators

- Bernoulli trials: $\hat{p}_n^{\mathrm{MLE}} = \bar{X}_n$.
- Poisson model: $\hat{\lambda}_n^{\mathrm{MLE}} = \bar{X}_n$.
- Gaussian model: $\left(\hat{\mu}_n, \hat{\sigma}_n^2\right)^{\mathrm{MLE}} = \left(\bar{X}_n, \hat{S}_n\right)$.
- Uniform model: $\hat{\theta}^{\mathrm{MLE}} = X_{(n)} = \max_i X_i$
- Mixture of Gaussians: no closed form. Need to use an optimization algorithm, for example EM.

The EM algorithm

To maximize the (log-) likelihood in mixtures of Gaussians, we often use the popular Expectation-Maximization (EM) algorithm.

- It is a *heuristic*. In particular, it can fail to find the MLE.
- Some very recent guarantees have been proved but require structural assumptions and/or good initialization.
- In practice, the algorithm is started from different random initializations and the solution with largest log-likelihood is kept in the end.
- The EM algorithm was introduced in 1977 and is still hugely popular

TITLE	CITED BY	YEAR
Maximum Likelihood from Incomplete Data Via the *EM* Algorithm AP Dempster, NM Laird, DB Rubin Journal of the Royal Statistical Society: Series B (Methodological) 39 (1), 1-22	61636	1977

Likelihood

To illustrate EM, assume that $\pi = 1/2$, and $\sigma_1^2 = \sigma_2^2 = 1$. Recall that the PDF is

$$f(x) = \frac{1}{2\sqrt{2\pi}} \left\{ e^{-\frac{(x-\mu_1)^2}{2}} + e^{-\frac{(x-\mu_2)^2}{2}} \right\}.$$

So log-likelihood is:

$$\ell(x_1, \ldots, x_m; \mu_1, \mu_2) = \sum_{i=1}^{n} \log \left[e^{-\frac{(x_i-\mu_1)^2}{2}} + e^{-\frac{(x_i-\mu_2)^2}{2}}) \right] - n \log(2\sqrt{2\pi})$$

Not easily tractable.

Complete observations

We also have the *sampling* description:

$$X = ZX^{(1)} + (1 - Z)X^{(2)}$$

Z is a *latent* variable with pmf $p(z) = \begin{cases} 1/2 & \text{if } z = 0 \\ 1/2 & \text{if } z = 1 \end{cases}$

What if we observed both (Z, X)? Their joint density is

$$\begin{aligned} f(x, z) &= p(z) \cdot f(x|z) \\ &= \frac{1}{2} \cdot \left(z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2}} + (1 - z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2}} \right) \\ &= \frac{1}{2\sqrt{2\pi}} e^{-z\frac{(x-\mu_1)^2}{2}} e^{-(1-z)\frac{(x-\mu_1)^2}{2}} \\ &= \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z(x-\mu_1)^2 + (1-z)(x-\mu_2)^2}{2} \right) \end{aligned}$$

Complete likelihood

The complete likelihood becomes

$$L^{\text{comp}}((x_1, z_1), \ldots, (x_n, z_n); \mu_1, \mu_2) = \prod_{i=1}^{n} \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z_i(x_i - \mu_1)^2 + (1 - z_i)(x_i - \mu_2)^2}{2} \right)$$

and the corresponding complete log-likelihood is

$$\ell^{\text{comp}}(\mu_1, \mu_2) = -n \log(2\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^{n} [Z_i(X_i - \mu_1)^2 + (1 - Z_i)(X_i - \mu_2)^2]$$

Easy to maximize:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n Z_i X_i}{\sum_{i=1}^n Z_i}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^n (1 - Z_i) X_i}{\sum_{i=1}^n (1 - Z_i)}$$

but requires knowledge of the Z_is which we don't have...

The E step

Idea: replace unknown Z_i by its (conditional) Expectation:

- First attempt: $Z_i \approx \mathbb{E}[Z_i] = 1/2$. This is **too rough**!
- Second attempt: $Z_i \approx \mathbb{E}[Z_i|X_i]$. This is much better!

$$\begin{aligned}\mathbb{E}[Z_i|X_i] &= \mathbb{P}[Z_i = 1|X_i] \\ &= \frac{f(X_i|Z_i=1)\mathbb{P}[Z_i=1]}{f(X_i|Z_i=1)\mathbb{P}[Z_i=1] + f(X_i|Z_i=0)\mathbb{P}[Z_i=0]} \\ &\qquad\qquad \text{(Bayes formula)} \\ &= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i-\mu_1)^2}{2}} \cdot \frac{1}{2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i-\mu_1)^2}{2}}\frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i-\mu_2)^2}{2}} \cdot \frac{1}{2}} \\ &= \frac{e^{-\frac{(X_i-\mu_1)^2}{2}}}{e^{-\frac{(X_i-\mu_1)^2}{2}} + e^{-\frac{(X_i-\mu_2)^2}{2}}} =: w_i \in (0,1)\end{aligned}$$

Note that w_i depends on μ_1, μ_2.

The M step

If we replace Z_i by $\mathbb{E}[Z_i|X_i] = w_i$ in the complete log-likelihood, we get

$$\tilde{\ell}(\mu_1, \mu_2) = -n\log(2\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}[w_i\ (X_i-\mu_1)^2 + (1-w_i)\ (X_i-\mu_2)^2]$$

Which is easy to maximize. It yields

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^n (1-w_i) X_i}{\sum_{i=1}^n (1-w_i)}$$

The EM algorithm

Input data: $X_1, \dots, X_n$.

1. Initialize $\hat{\mu}_1, \hat{\mu}_2$ (e.g. independent $\mathcal{N}(0,1)$)
2. Repeat until convergence:
   - Compute weights (E-step):

$$w_i \leftarrow \frac{e^{-\frac{(X_i-\mu_1)^2}{2}}}{e^{-\frac{(X_i-\mu_1)^2}{2}} + e^{-\frac{(X_i-\mu_2)^2}{2}}}, \quad i = 1, \dots, n$$

   - Update centers (M-step):

$$\hat{\mu}_1 \leftarrow \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \qquad \hat{\mu}_2 \leftarrow \frac{\sum_{i=1}^n (1-w_i) X_i}{\sum_{i=1}^n (1-w_i)}$$

Consistency of maximum likelihood estimator

Under mild regularity conditions, we have

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$$

This is because for all $\theta \in \Theta$

$$\frac{1}{n}\log L(X_1, \dots, X_n, \theta) \xrightarrow[n\to\infty]{\mathbb{P}} \text{“constant”} - \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$$

Moreover, the minimizer of the right-hand side is θ^* if the parameter is identifiable.

Technical conditions allow to transfer this convergence to the minimizers.

Fisher Information

Definition: Fisher information

Define the log-likelihood for one observation as:

$$\ell(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

$$I(\theta) = \mathsf{var}\big[\ell'(\theta)\big] = -\mathbb{E}\big[\ell''(\theta)\big]$$

Equivalence of the two definitions

We write it in the case of a continuous r.v. with pdf f_θ. It all starts with the $\star$ identity (we will write $\overset{\star}{=}$ when we use it):

$$\int f_\theta(x)dx = 1 \;\Rightarrow\; \frac{d}{d\theta}\int f_\theta(x)dx = \boxed{\int \frac{d}{d\theta} f_\theta(x)dx = 0} \qquad (\star)$$

We now compute $\mathsf{var}\big[\ell'(\theta)\big]$ and $-\mathbb{E}\big[\ell''(\theta)\big]$ and check that they are indeed equal. First we compute derivatives:

$$\ell'(\theta) = \frac{d}{d\theta}\log f_\theta(x) = \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}, \qquad \ell''(\theta) = \frac{\frac{d^2}{d\theta^2} f_\theta(x)}{f_\theta(x)} - \big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\big)^2.$$

The first identity gives

$$\begin{aligned}\mathsf{var}[\ell'(\theta)] = \mathbb{E}[(\ell'(\theta))^2] - (\mathbb{E}[\ell'(\theta)])^2 &= \int \big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\big)^2 f_\theta(x)dx - \big(\int \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x)dx\big)^2 \\ &= \int \frac{\big(\frac{d}{d\theta} f_\theta(x)\big)^2}{f_\theta(x)}dx - \big(\int \frac{d}{d\theta} f_\theta(x)dx\big)^2 \overset{\star}{=} \int \frac{\big(\frac{d}{d\theta} f_\theta(x)\big)^2}{f_\theta(x)}dx\end{aligned}$$

Moreover, the second identity gives

$$\begin{aligned}\mathbb{E}[\ell''(\theta)] &= \int \frac{\frac{d^2}{d\theta^2} f_\theta(x)}{f_\theta(x)} f_\theta(x)dx - \int \big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\big)^2 f_\theta(x)dx \\ &= \frac{d}{d\theta}\int \frac{d}{d\theta} f_\theta(x)dx - \int \frac{\big(\frac{d}{d\theta} f_\theta(x)\big)^2}{f_\theta(x)}dx \overset{\star}{=} -\int \frac{\big(\frac{d}{d\theta} f_\theta(x)\big)^2}{f_\theta(x)}dx = -\mathsf{var}[\ell'(\theta)].\end{aligned}$$

Fisher information of the Bernoulli experiment

Let $X \sim \mathsf{Ber}(p)$.

$$\ell(p) = \log(p^X(1-p)^(1-X)) = X\log p + (1-X)\log(1-p)$$

$$\ell'(p) = \frac{X}{p} - \frac{1-X}{1-p} \qquad \mathsf{var}[\ell'(p)] = \frac{1}{p(1-p)}$$

$$\ell''(p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} \qquad -\mathbb{E}[\ell''(p)] = \frac{1}{p(1-p)}$$

Asymptotic normality of the MLE

Theorem

Let $\theta^* \in \Theta$ (the *true* parameter). Assume the following:

1. The parameter is identifiable.
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ;
3. θ^* is not on the boundary of Θ;
4. $I(\theta) \neq 0$ in a neighborhood of θ^*;
5. A few more technical conditions.

Then, $\hat\theta_n^{\mathrm{MLE}}$ satisfies:

- $\hat\theta_n^{\mathrm{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- $\sqrt{n}\left(\hat\theta_n^{\mathrm{MLE}} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, I(\theta^*)^{-1}\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

An idea of the proof

We can use a technique resembling what we used for the Δ-method. How? We need to write the MLE as the function of an average. Write $\ell_i(\theta) := \log f_\theta(X_i)$ and by CLT, we have

$$\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}\ell_i'(\theta) - \mathbb{E}[\ell'(\theta)]\} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathsf{var}[\ell'(\theta)])$$

Note first that, $\mathbb{E}[\ell'(\theta)] \overset{\star}{=} 0$ and $\mathsf{var}[\ell'(\theta)] = I(\theta)$. Moreover, to make the MLE appear, recall that since it maximizes the log likelihood so that $\sum_{i=1}^{n}\ell_i'(\hat{\theta}^{\mathrm{MLE}}) = 0$.

Therefore, we can write we can write

$$\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}(\ell_i'(\hat{\theta}^{\mathrm{MLE}}) - \ell_i'(\theta)\} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta))$$

Now we start being more informal. Using a first order Taylor expansion (which is justified because the MLE is consistent), we have that

$$\frac{1}{\hat{\theta}^{\mathrm{MLE}} - \theta}\{\frac{1}{n}\sum_{i=1}^{n}(\ell_i'(\hat{\theta}^{\mathrm{MLE}}) - \ell_i'(\theta)\} \approx \{\frac{1}{n}\sum_{i=1}^{n}\ell_i''(\theta)\} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\ell_i''(\theta)] = -I(\theta) \quad \text{(LLN)}$$

The two above displays together with Slutsky yield

$$-I(\theta)\sqrt{n}(\hat{\theta}^{\mathrm{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta))$$

Dividing both sides by $-I(\theta)$ yields an asymptotic variance of $I(\theta)/I(\theta)^2 = 1/I(\theta)$.

M-estimation

MLE Strategy

Observe $X_1, \ldots, X_n \sim \mathbb{P}_{\theta^*}$, i.i.d, θ^* unknown.

1. Ideal loss function: $\theta \mapsto \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ minimized at $\theta = \theta^*$
2. Observe that $\mathsf{KL}(\mathbb{P}, \mathbb{P}_\theta) = -\mathbb{E}\log[p_\theta(X)]$ (plus additive constant)
3. Estimate by $-\frac{1}{n}\sum_{i=1}^{n}\log[p_\theta(X_i)]$ (-log-likelihood)
4. $\hat{\theta} := \mathrm{argmin}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log[p_\theta(X_i)]\right\}$

M-estimators

Idea:

- Let $X_1, \ldots, X_n$ be i.i.d with some unknown distribution $\mathbb{P}$ in some sample space E ($E \subseteq \mathbb{R}^d$ for some $d \geq 1$).
- No statistical model needs to be assumed (similar to ML).
- Goal: estimate some parameter μ^* associated with $\mathbb{P}$, e.g. its mean, variance, median, other quantiles, the true parameter in some statistical model...
- Find a function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values for the unknown μ^*, such that:

$$\mathcal{Q}(\mu) := \mathbb{E}\left[\rho(X_1, \mu)\right]$$

achieves its minimum at $\mu = \mu^*$.

Examples (1)

- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: $\mu^* =$
- If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x, \mu) = \|x - \mu\|_2^2$, for all $x \in \mathbb{R}^d, \mu \in \mathbb{R}^d$: $\mu^* =$
- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = |x - \mu|$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: μ^* is a median of $\mathbb{P}$.

Definition

- Define $\hat{\mu}_n$ as a minimizer of:

$$\mathcal{Q}_n(\mu) := \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \mu).$$

- Examples: Empirical mean, empirical median, empirical quantiles, MLE, etc.

MLE is an M-estimator

Assume that $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$ is a statistical model associated with the data.

Theorem

Let $\mathcal{M} = \Theta$ and $\rho(x, \theta) = -\log L_1(x, \theta)$, provided the likelihood is positive everywhere. Then,

$$\mu^* = \theta^*,$$

where $\mathbb{P} = \mathbb{P}_{\theta^*}$ (i.e., θ^* is the true value of the parameter).

The method of moments

Moments

Let X be a random variable with distribution $\mathbb{P}_\theta$ (write $\mathbb{E}_\theta$ for its expectation).

Definition

For $k = 1, 2, \ldots$, the **moment** of order k of X is given by

$$m_k = m_k(\theta) = \mathbb{E}_\theta[X^k]$$

Example 1: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned} m_1 &= \mathbb{E}[X] = \mu \\ m_2 &= \mathbb{E}[X^2] \\ &= \mathsf{var}[X] + (\mathbb{E}[X])^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Example 2: $X \sim \mathsf{Ber}(p)$

$$\begin{aligned} m_1 &= \mathbb{E}[X] &&= p \\ m_k &= \mathbb{E}[X^k] &&= p \end{aligned}$$

Moment generating function

For many distributions[2] $\mathbb{P}$, all the moments of $X \sim \mathbb{P}$ are contained in a *single* function called Moment Generating Function, or simply MGF:

$$M_X(t) = \mathbb{E}[e^{tX}] \qquad , t \in \mathbb{R}\,.$$

The moments are given by successive[3] derivatives of $M_X(\cdot)$ at $t = 0$:

$$\begin{aligned} M_X^{(1)}(t) &= \mathbb{E}\big[\tfrac{d}{dt} e^{tX}\big] &&= \mathbb{E}\big[Xe^{tX}\big] = \mathbb{E}[X] = m_1 \quad \text{for } t = 0 \\ M_X^{(2)}(t) &= \mathbb{E}\big[\tfrac{d^2}{dt^2} e^{tX}\big] &&= \mathbb{E}\big[X^2 e^{tX}\big] = \mathbb{E}[X^2] = m_2 \quad \text{for } t = 0 \\ &\vdots \\ M_X^{(k)}(t) &= \mathbb{E}\big[\tfrac{d^k}{dt^k} e^{tX}\big] &&= \mathbb{E}\big[X^k e^{tX}\big] = \mathbb{E}[X^k] = m_k \quad \text{for } t = 0 \end{aligned}$$

[2] It may be infinite for some t. For if X has a Cauchy distribution with pdf given by $f(x) = \frac{1}{\pi(1+x^2)}$

[3] For a function $f(t)$, we write $f^{(k)}(t) = \frac{d^k}{dt^k} f(t)$ for its kth derivative

MGF of a Standard Gaussian

Consider the Standard Gaussian r.v. $Z \sim \mathcal{N}(0, 1)$. We compute its MGF:

$$M_Z(t) = \mathbb{E}[e^{tZ}] = \frac{1}{\sqrt{2\pi}} \int e^{tz} e^{-\frac{z^2}{2}} dz$$

To compute it, we use a standard trick when manipulating Gaussians: *completing the square*

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int e^{tz} e^{-\frac{z^2}{2}} dz &= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-t)^2}{2}} e^{\frac{t^2}{2}} dz \\ &= e^{\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-t)^2}{2}} dz \\ &= e^{\frac{t^2}{2}} \cdot 1 \end{aligned}$$

Therefore

$$M_Z(t) = e^{\frac{t^2}{2}}$$

Moments of a Standard Gaussian

We have seen that for any r.v X,

$$m_k = M_X^{(k)}(0)\,, \quad k = 1, 2, \ldots$$

If $X = Z \sim \mathcal{N}(0, 1)$, compute

$$M_Z^{(k)}(0) = \frac{d^k}{dt^k} e^{\frac{t^2}{2}} \Big|_{t=0}$$

It yields

$$\begin{aligned} M_Z^{(1)}(t) &= te^{\frac{t^2}{2}} && \Rightarrow m_1 = M_Z^{(1)}(0) = 0 \\ M_Z^{(2)}(t) &= e^{\frac{t^2}{2}} + t^2 e^{\frac{t^2}{2}} && \Rightarrow m_2 = M_Z^{(2)}(0) = 1 \\ M_Z^{(3)}(0) &= 0 \\ M_Z^{(4)}(0) &= 3 \end{aligned}$$

Sample moments

- Statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$.
- Assume $\theta \subset \mathbb{R}^d$ (d parameters to estimate)
- Moments $m_k(\theta) = \mathbb{E}[X^k], k = 1, 2, \ldots$
- Let $X_1, \ldots, X_n$ be an i.i.d. observations from this model

The kth **sample moment** is

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k .$$

By LLN, we have

$$\hat{m}_k \xrightarrow[n\to\infty]{\mathbb{P}/a.s} m_k(\theta)$$

Recap

- Three principled methods for estimation: maximum likelihood, M-estimation, and the method of moments.
- Maximum likelihood is an example of M-estimation
- MLE tends to be best: asymptotic variance is smallest, given by inverse Fisher information.

Methods of moments estimator

Definition

The **methods of moments estimator** $\hat{\theta}_n \in \mathbb{R}^d$ satisfies

$$\begin{aligned} m_1(\hat{\theta}_n) &= \hat{m}_1 \\ m_2(\hat{\theta}_n) &= \hat{m}_2 \\ \vdots \quad &\quad \vdots \\ m_d(\hat{\theta}_n) &= \hat{m}_d \end{aligned}$$

This a system of d equations with d unknowns.

Ex. 1: $X \sim \mathcal{N}(\mu, \sigma^2)(d = 2)$ **Ex. 2**: $X \sim \text{Ber}(p)(d = 1)$

$$\begin{aligned} m_1 &= \mu \\ m_2 &= \sigma^2 + \mu^2 \end{aligned} \qquad m_1 = p$$

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = \left(\bar{X}_n, \bar{X}_n^2 - (\bar{X}_n)^2\right) \qquad \hat{p}_n = \bar{X}_n$$

18.650 – Fundamentals of Statistics

4. Parametric hypothesis testing

Waiting time in the ER

- The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- Collect a sample: $X_1, \dots, X_n$ (waiting time in minutes for n random patients) with unknown expected value $\mathbb{E}[X_1] = \mu$.
- We want to know if $\mu > 30$.

$$\begin{aligned} H_0 &: \quad \mu \leq 30 \\ H_1 &: \quad \mu > 30 \end{aligned}$$

Statistical formulation

- Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Let Θ_0 and Θ_1 be a *partition* of Θ.
- Consider the two hypotheses: $\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$
- H_0 is the *null hypothesis*, H_1 is the *alternative hypothesis*.
- We say that we *test* H_0 *against* H_1.

Testing lexicon

- For $k = 0$ (H_0) or $k = 1$ (H_1), we say that
  - Θ_k is a *simple hypothesis* if $\Theta_k = \{\theta_k\}$
  - Θ_k is a *composite hypothesis* if Θ_k is of the following three forms

$$\Theta_k = \{\theta : \ \theta > \theta_k\} \quad \Theta_k = \{\theta : \ \theta < \theta_k\} \quad \Theta_k = \{\theta : \ \theta \neq \theta_k\}$$

- A test is typically either *one-sided* or *two-sided*

Two-sided

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$$

One-sided

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_1 : & \theta > \theta_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0 : & \theta \geq \theta_0 \\ H_1 : & \theta < \theta_0 \end{cases}$$

Examples

1. Waiting time in the ER

$$\begin{aligned} H_0: &\quad \mu \leq 30 \\ H_1: &\quad \mu > 30 \end{aligned}$$

Both hypotheses are composite. The test is one-sided

2. In the Kiss example, we want to test

$$\begin{aligned} H_0: &\quad p = .5 \\ H_1: &\quad p \neq .5 \end{aligned}$$

H_0 is simple, H_1 is composite hypotheses. The test is two-sided

Clinical trials

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

Notation and modelling

- Let $\mu_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\mu_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.
- Hypothesis testing problem:

$$\begin{aligned} H_0: &\quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1: &\quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned}$$

- We observe two independent samples:
  - $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{d}}, \sigma_{\mathrm{d}}^2)$ from the test group and
  - $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_{\mathrm{c}}, \sigma_{\mathrm{c}}^2)$ from the control group.
- This is a *two-sample* test: these are very common (A/B testing).

Asymmetry in the hypotheses

- We want to decide whether to *reject* H_0 (look for evidence against H_0 in the data).
- H_0 and H_1 do not play a symmetric role: the data is only used to try to disprove H_0

$$\begin{aligned} H_0: &\quad \text{status quo} \\ H_1: &\quad \text{a (scientific) discovery} \end{aligned}$$

- In particular lack of evidence, does not mean that H_0 is true ("innocent until proven guilty")

Examples

1. Waiting time in the ER

$$\begin{aligned} H_0: &\quad \mu \leq 30 \\ H_1: &\quad \mu > 30 \end{aligned}$$

   Status quo: CDC statement. We collect data to show that Princeton-Plainsboro is different

2. Kiss

$$\begin{aligned} H_0: &\quad p = .5 \\ H_1: &\quad p \neq .5 \end{aligned}$$

   Status quo: our intuition tells us there should be no preference. We collect data to show that there is one.

3. Clinical trials

$$\begin{aligned} H_0: &\quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1: &\quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned}$$

   Status quo: The drug is not more effective than a placebo. We collect data to prove that the drug is effective.

Errors

A test can make two types of errors:

	Fail to reject Null	Reject Null
H_0 true ($\theta \in \Theta_0$)		
H_1 true ($\theta \in \Theta_1$)		

Both errors can be computed from the *power function*

$$\beta(\theta) = \mathbb{P}_\theta[\psi = 1]$$

- If $\theta \in \Theta_0$,

$$\beta(\theta) = \mathbb{P}_\theta[\psi \text{ makes an error of type I }]$$

  We want $\beta(\theta)$ to be small
- If $\theta \in \Theta_1$,

$$\beta(\theta) = 1 - \mathbb{P}_\theta[\psi \text{ makes an error of type II }]$$

  We want $\beta(\theta)$ to be large

What is a test?

- A *test* is a statistic $\psi \in \{0, 1\}$ that does not depend on unknown quantities and such that:
  - If $\psi = 0$, H_0 is not rejected;
  - If $\psi = 1$, H_0 is rejected.

  Important remark: Can always write $\psi = \mathbb{1}\{R\}$, where R is an *event* called *rejection region*

- Waiting time in the ER:

$$\begin{aligned} H_0: &\quad \mu \leq 30 \\ H_1: &\quad \mu > 30 \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n > C\}$$

- Kiss:

$$\begin{aligned} H_0: &\quad p = .5 \\ H_1: &\quad p \neq .5 \end{aligned} \qquad \psi = \mathbb{1}\{|\bar{X}_n - .5| > C\}$$

- Clinical trials

$$\begin{aligned} H_0: &\quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1: &\quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n - \bar{Y}_m > C\}$$

The Neyman-Pearson paradigm

Recall the waiting time in the ER example

$$\begin{aligned} H_0: &\quad \mu \leq 30 \\ H_1: &\quad \mu > 30 \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n > C\}$$

How to choose C ?

We are facing a dilemma: both errors should be small!

- To make Type I error $\to 0$, take $C \to +\infty$
- To make Type II error $\to 0$, take $C \to -\infty$

Cannot make both small at the same time.

The *Neyman-Pearson paradigm*:

- Make sure that $\mathbb{P}[\text{Type I error}] \leq \alpha$ (e.g., $\alpha = 5\%, 1\%, \dots$)
- Minimize $\mathbb{P}[\text{Type II error}]$ subject to this constraint

Level

The value of $\alpha \in (0,1)$ chosen in the Neyman-pearson paradigm is called level of a test

For which $\theta \in \Theta_0$ should we compute $\mathbb{P}_\theta[\psi = 1]$ (probability of Type I error)?

- A test ψ has *level* α if

$$\mathbb{P}_\theta[\psi = 1] \leq \alpha, \quad \forall\ \theta\ \in \Theta_0.$$
$$\iff \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi = 1] \leq \alpha$$

- A test $\psi = \psi_n$ has *asymptotic level* α if

$$\lim_{n\to\infty} \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi_n = 1] \leq \alpha,$$

A test for the Kiss example

We want to test:

$$H_0: \quad p = 0.5$$
$$H_1: \quad p \neq 0.5$$

We observe $R_1, \ldots, R_n \overset{iid}{\sim} \text{Ber}(p)$.

- Recall that

$$\mathcal{I}_{\text{conserv}} = \left[\bar{R}_n \pm \frac{1.96}{2\sqrt{n}}\right]$$

is a confidence interval of asymptotic level $1-\alpha$ for p.

- Consider the test:

$$\psi = \mathbb{1}\{.5 \notin \mathcal{I}_{\text{conserv}}\}$$

- We have

$$\lim_{n\to\infty} \mathbb{P}_{.5}[\psi = 1] = 1 - \lim_{n\to\infty} \mathbb{P}_{.5}\left[.5 \in \mathcal{I}_{\text{conserv}}\right] \leq 1 - (1-\alpha) = \alpha$$

- Therefore ψ is a test with asymptotic level α

Building a test from a confidence interval

Given a confidence interval, we can often build a test (and vice versa).

- Let $I = [A, B]$ be a confidence interval at level $1-\alpha$ for a parameter θ:

$$\mathbb{P}_\theta(\theta \in [A, B]) \geq 1 - \alpha$$

- We want to use this I to build a test at level α for

$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta \neq \theta_0$$

- Natural candidate:

$$\psi = \mathbb{1}\{\quad \theta_0 \notin [A, B] \quad\}$$

- Level of test:

$$\mathbb{P}_{\theta_0}[\psi = 1] = \mathbb{P}_{\theta_0}\left[\theta_0 \notin I\right] = 1 - \mathbb{P}_{\theta_0}\left[\theta_0 \in I\right] \leq 1 - (1-\alpha) = \alpha$$

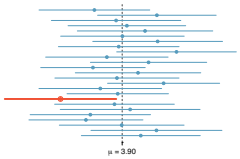
- Therefore ψ is a test with level α

Meaning of the level

- Recall that

$\mathcal{I}$ is a CI at level 95% for θ

means that if we repeat the experiment many times, at least 95% confidence intervals will contain the true parameter θ.



Confidence interval coverage plot, (c)OpenIntro.org

- Similarly:

ψ is a test at level 5% for H_0 vs H_1

means that if we repeat the experiment many times, at most 5% of the tests will make an error of type I.

What if we change the level?

With our data $\mathcal{I}_{\text{conserv}} = [0.56, 0.73]$ so we reject H_0 at level 5%.

α	$q_{\alpha/2}$	$\mathcal{I}_{\text{conserv}}$	decision
10%	1.64	$[0.57, 0.72]$	Reject
5%	1.96	$[0.56, 0.73]$	Reject
1%	2.76	$[0.52, 0.77]$	Reject
$.1\%$	3.29	$[0.497, 0.79]$	Fail to reject
$.01\%$	3.89	$[0.47, 0.82]$	Fail to reject

The value of α across which we switch from "reject" to "fail to reject" is called the p-value.

The evidence scale

- Statisticians, and more generally researchers, are used to communicating directly in terms of p-values rather than "reject/fail to reject at level..."
- The mental conversion is as follows:

p-value	evidence against H_0
$> 10\%$	almost none
$[5\%, 10\%]$	weak
$[1\%, 5\%]$	strong
$[.1\%, 1\%]$	very strong
$< .1\%$	undisputable

p-value

Definition

The (asymptotic) *p-value* of a test ψ is the smallest (asymptotic) level α at which ψ rejects H_0.

Golden rule

p-value $\leq \alpha \quad \Leftrightarrow \quad H_0$ is rejected by ψ, at the (asymptotic) level α.

Kiss example: we need to find α_0 such that $\bar{R}_n - \frac{q_{\alpha_0/2}}{2\sqrt{n}} = 0.5$

If $\bar{R}_n = .645,\ n = 124$ we get $q_{\alpha_0/2} = 3.23$. To find α_0:

$$\frac{\alpha_0}{2} = \mathbb{P}[Z > 3.23] = 1 - 0.9994 = 0.06\% \qquad \Rightarrow \ \alpha_0 = 0.12\%$$

where $Z \sim \mathcal{N}(0,1)$ and $\mathbb{P}(Z \leq 3.24) = 0.9994$ (read from table).

Parametric hypothesis testing

Parametric hypothesis testing

- Given the duality between confidence intervals (CI) and tests, it is not surprising that the same tools will be used.
- A simple approach: first build CI, then deduce a test is nice but *limited*: one/two-sided, two sample tests are more common than confidence intervals for (say) $\mu_{\mathrm{d}} > \mu_{\mathrm{c}}$.
- Easier to unfold the same machinery: this is the principle behind the Wald test.
- Wald's test only guarantees *asymptotic* level. An alternative is the T-test.

The Wald test (1)

- Statistical model $(E, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$
- Estimator $\hat{\theta}$ such that $\dfrac{\hat{\theta}-\theta}{\sqrt{\widehat{\mathsf{var}}(\hat{\theta})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$
  where $\widehat{\mathsf{var}}(\hat{\theta})$ is an estimator of the variance of $\hat{\theta}$
- For example, in the Bernoulli case, $\widehat{\mathsf{var}}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n}$.

The Wald test (2)

	$H_0: \ \theta = \theta_0$ $H_1: \ \theta \neq \theta_0$	$H_0: \ \theta \leq \theta_0$ $H_1: \ \theta > \theta_0$	$H_0: \ \theta \geq \theta_0$ $H_1: \ \theta < \theta_0$
Wald Test ψ	$\mathbb{1}(\lvert W\rvert > q_{\alpha/2})$	$\mathbb{1}(W > q_\alpha)$	$\mathbb{1}(W < -q_\alpha)$

$$W := \frac{\hat{\theta}-\theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat{\theta})}}$$

Asymptotic level of the Wald test (case 1)

If

$$\begin{aligned} H_0: &\ \theta = \theta_0 \\ H_1: &\ \theta \neq \theta_0 \end{aligned}$$

Then, for any $\theta = \theta_0$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}[\psi = 1] = \lim_{n\to\infty} \mathbb{P}_{\theta_0}[|W| > q_{\alpha/2}] = \mathbb{P}[|Z| > q_{\alpha/2}] = \alpha$$

Note that it is important to take the same θ_0 in $\mathbb{P}_{\theta_0}$ and W!

Asymptotic level of the Wald test (case 2 & 3)

If

$$\begin{aligned} H_0: &\quad \theta \leq \theta_0 \\ H_1: &\quad \theta > \theta_0 \end{aligned}$$

Then, for any $\theta \leq \theta_0$,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}_\theta[\psi = 1] &= \lim_{n\to\infty} \mathbb{P}_\theta[W > q_\alpha] \\ &= \lim_{n\to\infty} \mathbb{P}_\theta\left[\frac{\hat\theta - \theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat\theta)}} > q_\alpha\right] \\ &\leq \lim_{n\to\infty} \mathbb{P}_\theta\left[\frac{\hat\theta - \theta}{\sqrt{\widehat{\mathsf{var}}(\hat\theta)}} > q_\alpha\right] \\ &= \mathbb{P}[Z > q_\alpha] = \alpha \end{aligned}$$

p-values for the Wald test

- Recall that $W := \dfrac{\hat\theta - \theta_0}{\sqrt{\widehat{\mathsf{var}}(\hat\theta)}}$.
- Denote by W^{obs} the realization (observed value) of W in a given example. For the News example, $W^{\text{obs}} = 1.93$
- Then p-values and asymptotic p-values are given by

	$H_0: \ \theta = \theta_0$ $H_1: \ \theta \neq \theta_0$	$H_0: \ \theta \leq \theta_0$ $H_1: \ \theta > \theta_0$	$H_0: \ \theta \geq \theta_0$ $H_1: \ \theta < \theta_0$
Wald test	$\lvert W\rvert > q_{\alpha/2}$	$W > q_\alpha$	$W < -q_\alpha$
p-value	$\mathbb{P}(\lvert W\rvert > \lvert W^{\text{obs}}\rvert)$	$\mathbb{P}(W > W^{\text{obs}})$	$\mathbb{P}(W < W^{\text{obs}})$
asymp. p-value	$\mathbb{P}(\lvert Z\rvert > \lvert W^{\text{obs}}\rvert)$	$\mathbb{P}(Z > W^{\text{obs}})$	$\mathbb{P}(Z < W^{\text{obs}})$

where $Z \sim \mathcal{N}(0,1)$

Example 1: News

More than 2/3 of Americans get news on social media

Is this quote from a 2018 Pew Research Center study justified?
$X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p)$, $p \in [0,1]$,

$$\begin{aligned} H_0: &\quad p \leq 2/3 \\ H_1: &\quad p > 2/3 \end{aligned}$$

This claim is based on $n = 4,581$ randomly sampled U.S., $\hat p = .68$.

$$W^{\text{obs}} = \sqrt{4,581}\frac{.68 - 2/3}{\sqrt{.68(1-.68)}} = 1.93 \quad > 1.645 \text{ so Reject}$$

The p-value is α_0 such that

$$q_{\alpha_0} = 1.93 \iff \alpha_0 = \mathbb{P}(Z > 1.93) = 1 - 0.9732 = 2.68\%$$

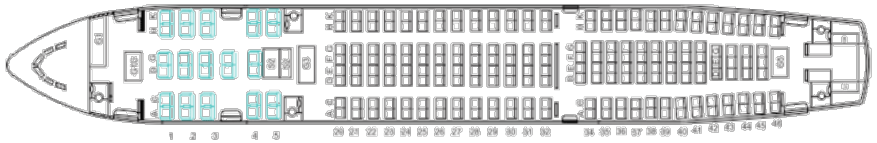
Fail to reject at asymptotic level $\alpha = 1\%$.

Example 2: How to board a plane?

What is the fastest method to board a plane?

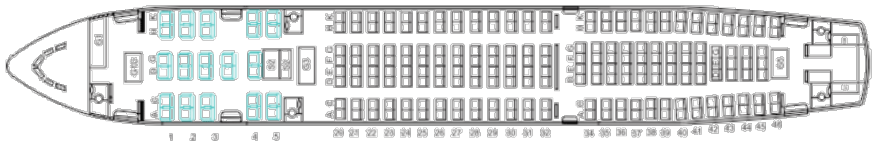
R2F or WilMA?

- R2F= Rear to Front (JetBlue)



(c) airbus.com

- WilMA=Window, Middle, Aisle (United)



(c) airbus.com

Model and Assumptions

- X: boarding time of a random JetBlue flight.

$$\mathbb{E}[X] = \mu_1, \qquad \text{var}[X] = \sigma_1^2$$

- Y: boarding time of a random United flight.

$$\mathbb{E}[Y] = \mu_2, \qquad \text{var}[Y] = \sigma_2^2$$

- We have $X_1, \ldots, X_n$ independent copies of X and $Y_1, \ldots Y_m$ independent copies of Y.
- We further assume that the two samples are independent.

Is there a difference between the two boarding methods:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned}$$

Equivalently, write $\theta = \mu_1 - \mu_2$, we get

$$\begin{aligned} H_0 &: \quad \theta = 0 \\ H_1 &: \quad \theta \neq 0 \end{aligned}$$

We have two samples: this is a **two-sample** testing problem.

Asymptotically normal estimator for θ

- Define the estimator $\hat{\theta} = \bar{X}_n - \bar{Y}_m$.
- We have by the CLT:

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{var}(\hat{\theta})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

- But: $\text{var}(\hat{\theta}) = \text{var}(\bar{X}_n) + \text{var}(\bar{Y}_m) = \dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}$
- We can estimate σ_1^2 by $\hat{\sigma}_1^2$ and σ_2^2 by $\hat{\sigma}_2^2$ where

$$\hat{\sigma}_1^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad \hat{\sigma}_2^2 := \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$$

- Both estimators are consistent so by Slutsky

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \xrightarrow[\substack{n\to\infty \\ m\to\infty}]{(d)} \mathcal{N}(0,1)$$

Applying the Wald test

$$W = \frac{\hat{\theta} - 0}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \qquad \psi = \mathbb{1}\{|W| > q_{\alpha/2}\}$$

Data from JetBlue (R2F) and United (WilMA):

	R2F	WilMA
Average (mins)	24.2	25.9
Std. Dev (mins)	5.1	4.3
Sample size	72	56

$$W = \frac{24.2 - 25.9}{\sqrt{\frac{5.1^2}{72} + \frac{4.3^2}{56}}} = -2.04$$

The (asymptotic) p-value is given by

$$\alpha_0 = \mathbb{P}[|Z| > 2.04] = 2\mathbb{P}[Z < -2.04] = 2 \cdot 0.0207 = 4.14\%$$

Example 3: Waiting for the T

Waiting times for the T: $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$.

$$\begin{aligned} H_0 &: \quad \lambda \geq 1 \\ H_1 &: \quad \lambda < 1 \end{aligned}$$

- Recall that using the Delta-method, we got for $\hat{\lambda} = 1/\bar{X}_n$,

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \lambda^2)$$

- Therefore, by Slutsky

$$\sqrt{n}\frac{1}{\hat{\lambda}}(\hat{\lambda} - \lambda) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

- Test statistic

$$W = \sqrt{n}\frac{1}{\hat{\lambda}}(\hat{\lambda} - 1)$$

- Reject at 5% if $W < 1.645$.

Example 4: MLE and the Wald test

- Recall that under some regularity conditions, we have:

$$\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \frac{1}{I(\theta)})$$

  where $I(\theta)$ is the Fisher information.
- Using Slutsky, we get

$$\sqrt{nI(\hat{\theta}^{\text{MLE}})}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1)$$

- Therefore, we can use the Wald test with test statistic given by

$$W = \sqrt{nI(\hat{\theta}^{\text{MLE}})}(\hat{\theta}^{\text{MLE}} - \theta_0)$$

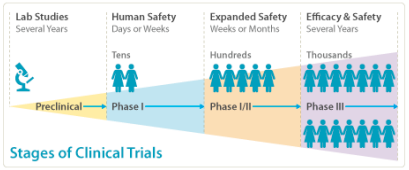
Who was Wald?

- Abraham Wald (1902-1950) was a very influential statistician and mathematician
- First (correct) proof of consistency of MLE under general conditions
- Introduced the first notion of curvature of a metric space
- Worked on aircraft damage during WWII
- Died in a plane crash in India while giving lectures across the country

Photo by onrad acobs, Erlangen, (c) he Mathematisches orschungsinstitut Oberwolfach gGmb (M O). CC B -SA 2.0 Germany

Small sample sizes

- Sometimes, sample sizes are too small to apply CLT/Slutsky which is central to the Wald test.
- This is often the case for early phases clinical trials
- No magic: we have to assume that our data is **Gaussian**



Home wind turbines

- The DoE recommends a minimum average wind speed of 10 miles an hour for a grid-connected wind turbine
- A candidate home was monitored once a month for a year and 12 measurements $X_1, \ldots, X_{12}$ were collected
- Assume that

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- Can we conclude that there is enough wind at this home?

Photograph © 2022 by Rob Cardillo

What goes wrong?

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- We don't even need the CLT since $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$
- But to replace σ^2 with a consistent estimator $\hat{\sigma}^2$, we need Slutsky, which is not true for small sample sizes:
- We carefully choose an unbiased estimator $\hat{\sigma}^2$:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- We are going to use the representation:

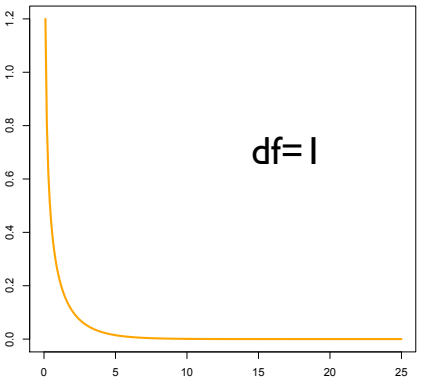
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} = \frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}}$$

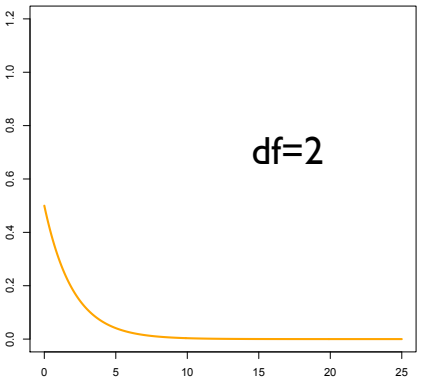
The χ^2 distribution

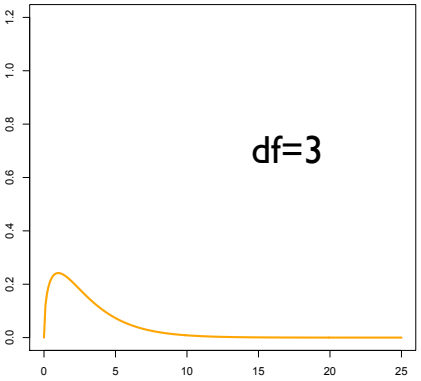
Definition

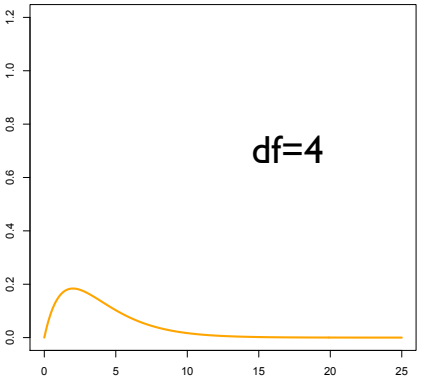
For a positive integer k, the χ^2 *(pronounced "Kai-squared") distribution with k degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_k^2$, where $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$.

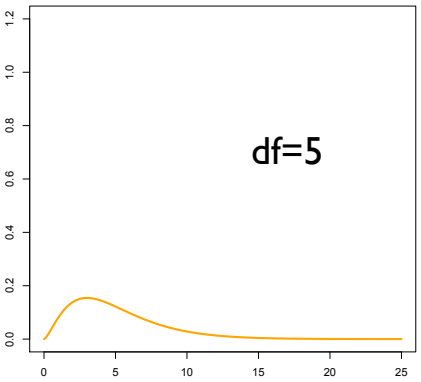
Example: If $Z \sim \mathcal{N}_k(\mathbf{0}, I_k)$, then $\|Z\|_2^2 \sim \chi_k^2$.

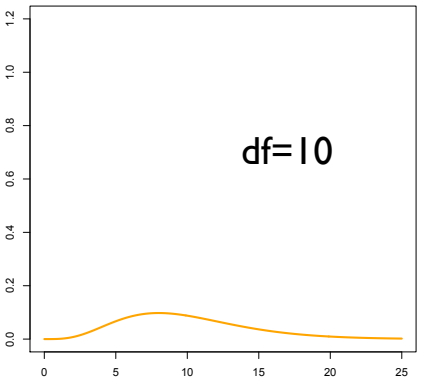


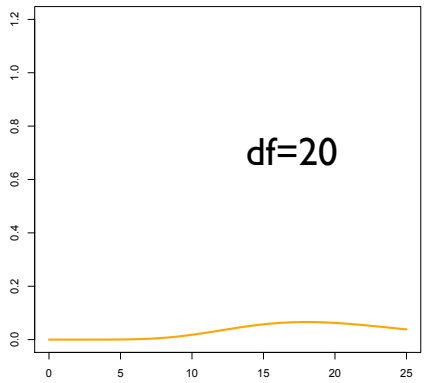



df=3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=10
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25



Properties of the χ^2 distribution

Definition
For a positive integer k, the χ^2 *(pronounced "Kai-squared") distribution with k degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_k^2$, where $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$.

Properties: If $V \sim \chi^2_k$, then

- $\mathbb{E}[V] = k$
- $\text{var}[V] = 2k$

Important example: the sample variance

- Recall that $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$
- **Cochran's theorem:** If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then
  - $\dfrac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\dfrac{X_i - \bar{X}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}$.
  - $\bar{X}_n$ and S_n^2 are independent r.v.;
- Therefore

$$\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}}$$

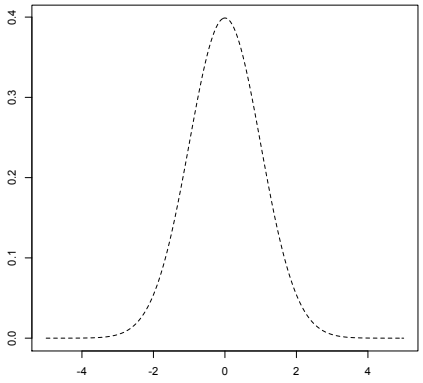
where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent

Student's T distribution

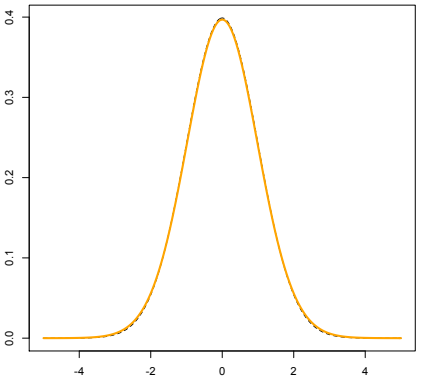
Definition
For a positive integer k, *Student's T distribution with k degrees of freedom* (denoted by t_k) is the law of the random variable

$$\frac{Z}{\sqrt{V/k}}$$

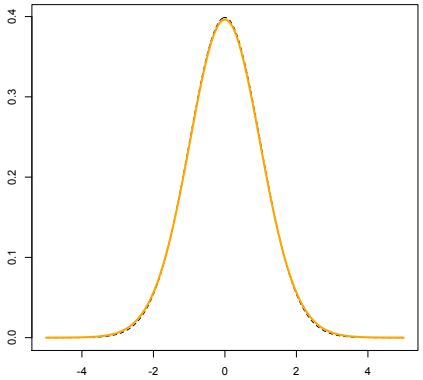
where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_k$ are independent r.v.



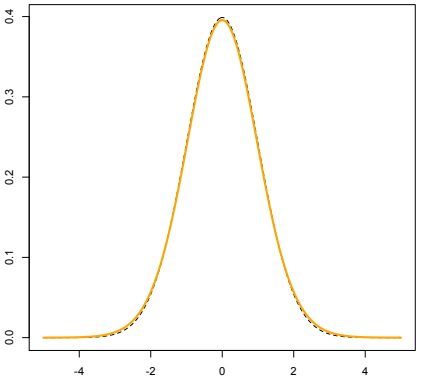
The standard normal pdf



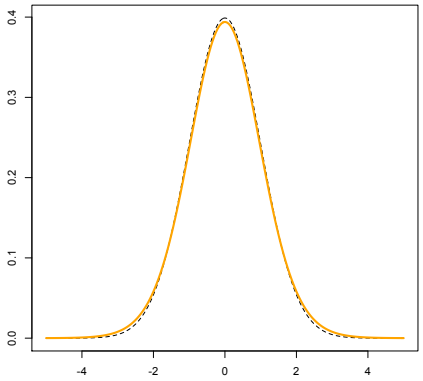
The t_{50} pdf



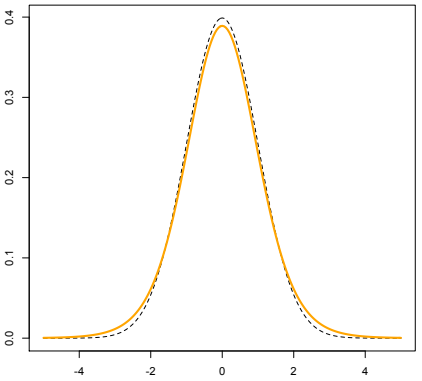
The t_{40} pdf



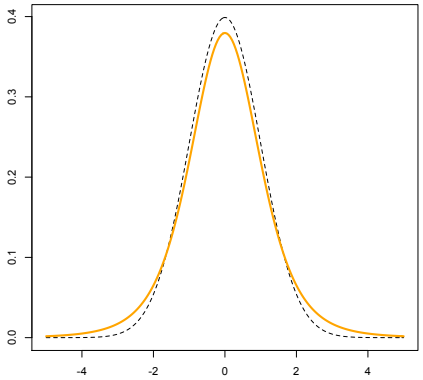
The t_{30} pdf



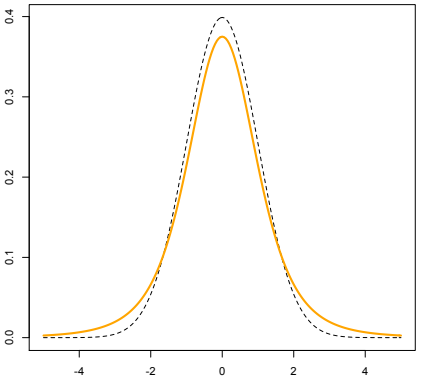
The t_{20} pdf



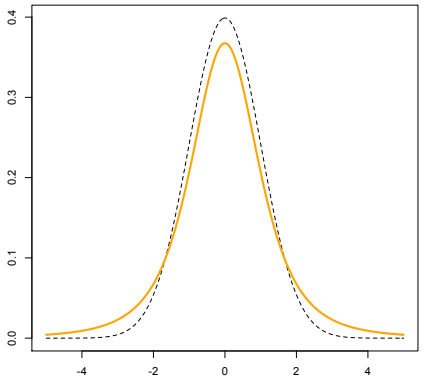
The t_{10} pdf



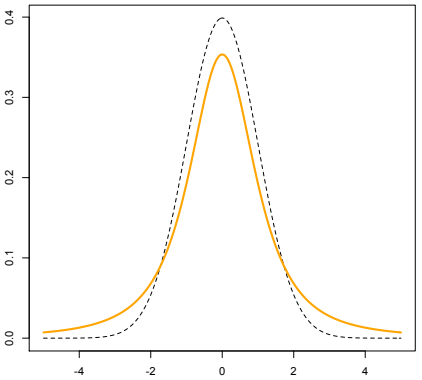
The t_5 pdf



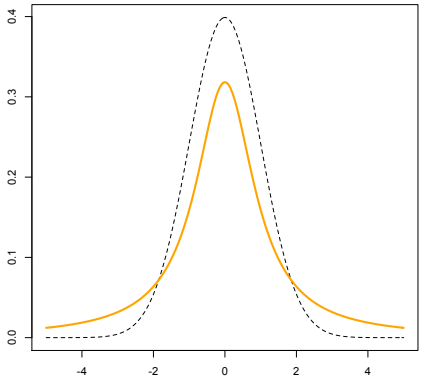
The t_4 pdf



The t_3 pdf



The t_2 pdf



The t_1 pdf

Student's T test (one sample)

- Gaussian model: $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ both μ and σ^2 unknown
- We know that
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}$$
- Then T-tests are given by:

	$H_0: \ \mu = \mu_0$ $H_1: \ \mu \neq \mu_0$	$H_0: \ \mu \leq \mu_0$ $H_1: \ \mu > \mu_0$	$H_0: \ \mu \geq \mu_0$ $H_1: \ \mu < \mu_0$
T-Test ψ	$\mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{1}\{T > q_{\alpha}^{t_{n-1}}\}$	$\mathbb{1}\{T < -q_{\alpha}^{t_{n-1}}\}$

where
$$T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n} \qquad \text{and} \qquad \mathbb{P}[t_{n-1} > q_{\alpha}^{t_{n-1}}] = \alpha$$

- Using the same analysis as for Wald's test, it can be shown that the T-test has (non-asymptotic) level α

Example: Home wind turbines

- We observe: $n = 12, \quad \bar{X}_n = 14.3 \quad S_n = 4.7 \quad \mu_0 = 10$

$$\begin{aligned} H_0 &: \quad \mu \leq 10 \\ H_1 &: \quad \mu > 10 \end{aligned}$$

- We get $T = \sqrt{12}\dfrac{14.3 - 10}{4.7} = 3.17$
- From table: $q_{5\%}^{t_{11}} = 1.80$ so we reject since $3.17 > 1.80$

p-values for the T-test

	$\begin{aligned} H_0 &: \ \mu = \mu_0 \\ H_1 &: \ \mu \neq \mu_0 \end{aligned}$	$\begin{aligned} H_0 &: \ \mu \leq \mu_0 \\ H_1 &: \ \mu > \mu_0 \end{aligned}$	$\begin{aligned} H_0 &: \ \mu \geq \mu_0 \\ H_1 &: \ \mu < \mu_0 \end{aligned}$
T-Test	$\mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{1}\{T > q_{\alpha}^{t_{n-1}}\}$	$\mathbb{1}\{T < -q_{\alpha}^{t_{n-1}}\}$
p-value	$\mathbb{P}[\lvert T\rvert > \lvert T^{\text{obs}}\rvert]$	$\mathbb{P}[T > T^{\text{obs}}]$	$\mathbb{P}[T < T^{\text{obs}}]$

where $T \sim t_{n-1}$

Home wind turbines: p-value=$\mathbb{P}[t_{11} > 3.17] = 0.446\%$ (very strong evidence)

Comparison with the Wald test

- If this was a Wald test at asymptotic level 5%, we would compare this to $q_{5\%} = 1.645$ and reject
- In this setup, only difference is $q_{5\%}^{t_{11}} = 1.80$ vs. $q_{5\%} = 1.645$ (Gaussian quantiles).
- For n large enough, difference becomes very small.
- But in general, Wald test is more flexible (any θ, $\hat{\text{var}}(\hat{\theta})$, etc.)

α	10%	5%	2.5%	1%	0.5%
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
⋮					
30	1.31	1.70	2.04	2.46	2.75
⋮					
50	1.30	1.68	2.01	2.40	2.68
∞	1.28	1.65	1.96	2.33	2.58

Clinical trial example (1)

- $\mu_{\text{d}} > 0$: expected decrease of LDL in test group.
- $\mu_{\text{c}} > 0$: expected decrease of LDL level in control group.
- Hypothesis testing problem:

$$\begin{aligned} H_0 &: \quad \mu_{\text{d}} - \mu_{\text{c}} \leq 0 \\ H_1 &: \quad \mu_{\text{d}} - \mu_{\text{c}} > 0 \end{aligned}$$

- We observe two independent samples:
  - $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_{\text{d}}, \sigma_{\text{d}}^2)$ from the test group and
  - $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_{\text{c}}, \sigma_{\text{c}}^2)$ from the control group.
- We have

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_{\text{d}} - \mu_{\text{c}})}{\sqrt{\frac{\sigma_{\text{d}}^2}{n} + \frac{\sigma_{\text{c}}^2}{m}}} \sim \mathcal{N}(0, 1)$$

Clinical trial example (2)

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_\mathrm{d} - \mu_\mathrm{c})}{\sqrt{\frac{S_\mathrm{d}^2}{n} + \frac{S_\mathrm{c}^2}{m}}} = \frac{\frac{\bar{X}_n - \bar{Y}_m - (\mu_\mathrm{d} - \mu_\mathrm{c})}{\sqrt{\frac{\sigma_\mathrm{d}^2}{n} + \frac{\sigma_\mathrm{c}^2}{m}}}}{\sqrt{\frac{\frac{S_\mathrm{d}^2}{n} + \frac{S_\mathrm{c}^2}{m}}{\frac{\sigma_\mathrm{d}^2}{n} + \frac{\sigma_\mathrm{c}^2}{m}}}}$$

- Good news: $\dfrac{\bar{X}_n - \bar{Y}_m - (\mu_\mathrm{d} - \mu_\mathrm{c})}{\sqrt{\frac{\sigma_\mathrm{d}^2}{n} + \frac{\sigma_\mathrm{c}^2}{m}}} \sim \mathcal{N}(0, 1)$
- Bad news: don't know the distribution of $\sqrt{\dfrac{\frac{S_\mathrm{d}^2}{n} + \frac{S_\mathrm{c}^2}{m}}{\frac{\sigma_\mathrm{d}^2}{n} + \frac{\sigma_\mathrm{c}^2}{m}}}$

Two-sample T-test

- We have approximately

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_\mathrm{d} - \mu_\mathrm{c})}{\sqrt{\frac{S_\mathrm{d}^2}{n} + \frac{S_\mathrm{c}^2}{m}}} \sim t_N$$

  where

$$N = \frac{\left(S_\mathrm{d}^2/n + S_\mathrm{c}^2/m\right)^2}{\frac{S_\mathrm{d}^4}{n^2(n-1)} + \frac{S_\mathrm{c}^4}{m^2(m-1)}} \gtrsim \min(n, m)$$

- This is Welch-Satterthwaite (WS) formula
- Sanity check: if $m \to \infty$ (one sample limit), we have $N \to n - 1$.

Non-asymptotic test

- Example $n = 12, m = 22$, $\bar{X}_n = 156.4$, $\bar{Y}_m = 132.7$, $S_\mathrm{d} = 22.5$, $S_\mathrm{c} = 8.7$,

$$T = \frac{156.4 - 132.7}{\sqrt{\frac{22.5^2}{12} + \frac{8.7^2}{22}}} = 3.51$$

- Using the WS formula:

$$N = \frac{\left(22.5^2/12 + 8.7^2/22\right)^2}{\frac{22.5^4}{12^2 \cdot 11} + \frac{8.7^4}{22^2 \cdot 21}} = 12.82$$

  we round to 12.
- We get

$$\text{p-value} = \mathbb{P}(t_{12} > 3.51) \quad = 0.215\%$$

  So we reject the test: there is strong evidence that the drug is effective.

Who was Student?

William Sealy Gosset.
Image Credit: Wikipedia.org

This distribution was introduced by **William Sealy Gosset** (1876–1937) in 1908 while he was "head experimental brewer" for the Guinness brewery in Dublin, Ireland. He published his work under the pseudonym "Student" because Guinness forbid its employees to publish their results.

VOLUME VI MARCH, 1908 NO. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

Discussion

Advantage of Student's test: Non asymptotic / Can be run on small samples

Drawback of Student's test: It relies on the assumption that the sample is Gaussian (Next unit: we will see how to test this assumption)

The dead salmon experiment: setup

image (C) Craig Bennett (c) prefrontal.org., All rights reserved.

In 2009, neuroscientist Craig Bennett purchased a whole Atlantic salmon, took it to a lab at Dartmouth, and put it into an fMRI machine used to study the brain.

"The salmon was shown a series of $n = 15$ photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."

The dead salmon experiment:results

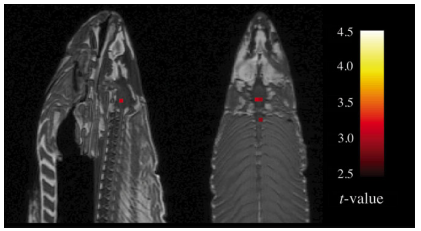


image (c) Craig Bennett (c) prefrontal.org., All rights reserved

- The salmon brain was split into $N = 8,064$ voxels
- In each voxel a statistical test at level $\alpha = .1\%$ was performed to see if activity was statistically significant
- 8 voxels* were found to be.
- Conclusion: these voxels contain salmon neurons involved in social perception. NO!

Statistical analysis

- Statistical model for a fixed voxel: observe (normalized) signal $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$, $n = 15$.
- Hypothesis testing problem:

$$\begin{aligned} H_0 &: \quad \mu = 0 \\ H_1 &: \quad \mu \neq 0 \end{aligned}$$

- Apply the T-test:
- Recall the interpretation of level $\alpha = .1\%$ of a test:

  If we repeat the experiment many times, at most .1% of the tests will make an error of type I.

- $.1\% \cdot 8,064 \simeq 8$

Multiple testing

- We cannot make conclusions about "all the voxels" at once: we are bound to make mistakes
- Two solutions:
  - Control Family Wise Error Rate (FWER): Find $C_1, \dots, C_N$ such that

$$\mathbb{P}_{\mu_i=0}\Big(\bigcup_{i=1}^{N}\{|T_i| > C_i\}\Big) \leq \alpha$$

  - Control False Discovery Rate (FDR): Find $C_1, \dots, C_N$ such that

$$\text{FDR} = \mathbb{E}\Big[\frac{\#\{i \,:\, |T_i| > C_i \;\&\; \mu_i = 0\}}{\#\{i \,:\, |T_i| > C_i\}}\Big] = \mathbb{E}_{\mu=0}\Big[\frac{\#\text{ of False discoveries}}{\#\text{ of discoveries}}\Big] \leq \alpha$$

- In both cases, it is easier to work with the p-values:

$$P_i = \mathbb{P}_{\mu_i=0}(|T| > |t_i^{\text{obs}}|)$$

where t_i^{obs} is the observed value of the test statistic for the ith test and $T \sim t_{n-1}$.

The Bonferroni method

- To control FWER, we use use the *Bonferroni correction*.
- Rather than rejecting each test at level α, we use the (much smaller) level α/N.
- In other words:

$$\text{Reject } i\text{th test} \quad \Longleftrightarrow \quad P_i < \frac{\alpha}{N}$$

- In the salmon example this means that each test is performed at level $0.001/8064 = 1.24 \cdot 10^{-7}$
- Often this is way to *conservative* (no discoveries).

Note that, with the Bonferroni method:

$$\begin{aligned} \text{FWER} &= \mathbb{P}_{\mu_i=0}\Big(\bigcup_{i=1}^{N}\{P_i < \frac{\alpha}{N}\}\Big) = \mathbb{P}_{\mu_i=0}\Big(\bigcup_{i=1}^{N}\{|T_i| > q_{\frac{\alpha}{2N}}\}\Big) \\ &\leq \sum_{i=1}^{N} \mathbb{P}_{\mu_i=0}\big(\{|T_i| > q^{t_{n-1}}_{\frac{\alpha}{2N}}\}\big) = N \cdot \frac{\alpha}{N} = \alpha \end{aligned}$$

The Benjamini-Hochberg method

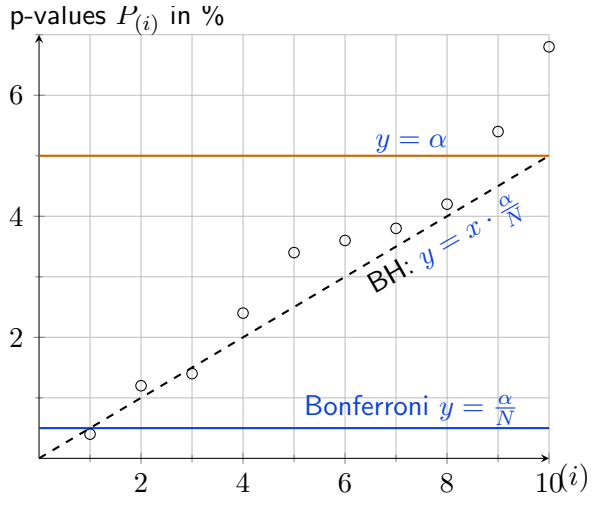
- To control FDR, we use use the *Benjamini-Hochberg (BH)* method.
- Intuitively, we should reject tests with the smallest p-values
- Order p-values: $P_{(1)} < P_{(2)} < P_{(3)} < P_{(4)} < \cdots < P_{(N)}$ and call "the (i)th test" the test with p-value $P_{(i)}$.
- Idea: reject all tests (i) such that $i \leq i_{\max}$
- Rule

$$i_{\max} := \max\Big\{i \,:\, P_{(i)} < i \cdot \frac{\alpha}{N}\Big\}$$

- Benjamini and Hochberg (1995, 68K citations) have shown that with this procedure:

$$\text{FDR} \leq \alpha$$

- There are *many* variations of the BH procedure, in particular to account for correlations between p-values.



In this figure, $N = 10, \alpha = 5\%$.

- Bonferonni: $\alpha/N = .005$. Only test (1) is rejected.
- BH: $i_{\max} = 3$. Tests $(1), (2)$ and (3) are rejected.

Recap

- Given an asymptotically normal estimator, we can build the Wald test using quantiles of the Gaussian
- When the data is Gaussian, we can use the T-test even for small sample sizes (quantiles of Student's T distribution)
- For large sample sizes, the quantiles of Student's T distribution converge to those of the Gaussian distribution
- When performing multiple tests, one needs to be careful and apply a correction: Bonferroni controls the FWER but is very conservative, BH controls the FDR and is very popular because it is less conservative.
- The statements FWER $\leq \alpha$ and FDR $\leq \alpha$ are often erroneously conflated but they have different meanings that

18.650 – Fundamentals of Statistics

5. Nonparametric hypothesis testing

Goodness of fit tests

Let X be a r.v. Given i.i.d copies of X we want to answer the following types of questions:

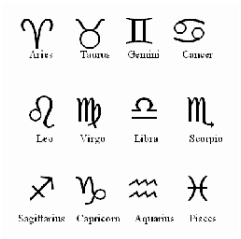
- Does X have distribution $\mathcal{N}(0,1)$? (Cf. Student's T distribution)
- Does X have distribution $\mathcal{U}([0,1])$?
- Does X have PMF $p_1 = 0.3,\ p_2 = 0.5,\ p_3 = 0.2$

These are all *goodness of fit* (GoF) tests: we want to know if the hypothesized distribution is a good fit for the data.

Key characteristic of GoF tests: no parametric modeling.

The zodiac sign of the most powerful people is....

Can your zodiac sign predict how successful you will be later in life? Fortune magazine collected the signs of 256 heads of the Fortune 500.



Fyi: 256/12 =21.33

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

The zodiac sign of the most successful people is....

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

In view of this data, is there statistical evidence that successful people are more likely to be born under some sign than others?

275 jurors with identified racial group.
We want to know if the jury is representative of the population of this county.

Race	White	Black	Hispanic	Other	Total
# jurors	205	26	25	19	275
proportion in county	0.72	0.07	0.12	0.09	1

Goodness of fit test

- Let $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_{\mathbf{p}}$, for some unknown $\mathbf{p} \in \Delta_K$, and let $\mathbf{p}^0 \in \Delta_K$ be fixed.
- We want to test:

$$H_0\colon\ \mathbf{p} = \mathbf{p}^0 \text{ vs. } H_1\colon\ \mathbf{p} \neq \mathbf{p}^0$$

with asymptotic level $\alpha \in (0,1)$.

- Example: If $\mathbf{p}^0 = (1/K, 1/K, \dots, 1/K)$, we are testing whether $\mathbb{P}_{\mathbf{p}}$ is the uniform distribution on E.

Discrete distribution

Let $E = \{a_1, \dots, a_K\}$ be a finite space and $(\mathbb{P}_{\mathbf{p}})_{\mathbf{p} \in \Delta_K}$ be the family of all probability distributions on E:

- $\Delta_K = \left\{ \mathbf{p} = (p_1, \dots, p_K) \in (0,1)^K : \sum_{j=1}^K p_j = 1 \right\}$.
- For $\mathbf{p} \in \Delta_K$ and $X \sim \mathbb{P}_{\mathbf{p}}$,

$$\mathbb{P}_{\mathbf{p}}[X = a_j] = p_j, \quad j = 1, \dots, K.$$

PMF, likelihood and maximum likelihood estimator

- Let $X \in \{a_1, \dots, a_K\}$ have pmf

$$p(a_j) = \mathbb{P}[X = a_j] = p_j, \quad j = 1, \dots, K$$

We can write

$$p(x) = \prod_{j=1}^K p_j^{\mathbb{1}(x = a_j)}$$

- Likelihood of the model:

$$L_n(X_1, \dots, X_n, \mathbf{p}) = p_1^{N_1} p_2^{N_2} \dots p_K^{N_K},$$

where $N_j = \#\{i = 1, \dots, n : X_i = a_j\}$.

- Let $\hat{\mathbf{p}}$ be the MLE: $\hat{\mathbf{p}}_j = \dfrac{N_j}{n}, \quad j = 1, \dots, K.$

⚠ $\hat{\mathbf{p}}$ maximizes $\log L_n(X_1, \dots, X_n, \mathbf{p})$ **under the constraint**

$$\sum_{j=1}^K p_j = 1.$$

χ^2 test

Theorem

$$\underbrace{n \sum_{j=1}^{K} \frac{\left(\hat{\mathbf{p}}_j - \mathbf{p}_j^0\right)^2}{\mathbf{p}_j^0}}_{T_n} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-1}.$$

- χ^2 test with asymptotic level α: $\psi = \mathbb{1}\{T_n > q_\alpha^{\chi^2_{K-1}}\}$, where $q_\alpha^{\chi^2_{K-1}}$ is the $(1-\alpha)$-quantile of χ^2_{K-1}.
- (Asymptotic) p-value of this test: $p - \text{value} = \mathbb{P}\left[Z > T_n^{\text{obs}}\right]$, where $Z \sim \chi^2_{K-1}$

CDF and empirical CDF

Let $X_1, \dots, X_n$ be i.i.d. real random variables. Recall the cdf of X_1 is defined as:

$$F(t) = \mathbb{P}[X_1 \leq t], \quad \forall t \in \mathbb{R}.$$

It completely characterizes the distribution of X_1.

Definition

The *empirical cdf* of the sample $X_1, \dots, X_n$ is defined as:

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} \\ &= \frac{\#\{i = 1, \dots, n : X_i \leq t\}}{n}, \quad \forall t \in \mathbb{R}. \end{aligned}$$

Consistency

By the LLN, for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (*Fundamental theorem of statistics*)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Asymptotic normality

By the CLT, for all $t \in \mathbb{R}$,

$$\sqrt{n}\ (F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\big(0, F(t)\,(1 - F(t))\big).$$

Donsker's Theorem

If F is continuous, then

$$\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0\leq t\leq 1} |\mathbb{B}(t)|,$$

where $\mathbb{B}$ is a Brownian bridge on $[0, 1]$.

Goodness of fit for continuous distributions

- Let $X_1, \dots, X_n$ be i.i.d. real random variables with unknown cdf F and let F^0 be a **continuous** cdf.
- Consider the two hypotheses:

$$H_0 : F = F^0 \quad \text{v.s.} \quad H_1 : F \neq F^0.$$

- Let F_n be the empirical cdf of the sample $X_1, \dots, X_n$.
- If $F = F^0$, then $F_n(t) \approx F^0(t)$, for all $t \in [0, 1]$.

Kolmogorov-Smirnov test

- Let $T_n = \sup_{t \in \mathbb{R}} \left| F_n(t) - F^0(t) \right|$.
- By Donsker's theorem, if H_0 is true, then $\sqrt{n} T_n \xrightarrow[n \to \infty]{(d)} Z$, where Z has a known distribution (supremum of a Brownian bridge).
- **KS test with asymptotic level α:**

$$\delta_\alpha^{KS} = \mathbb{1}\{T_n > q_\alpha / \sqrt{n}\},$$

where q_α is the $(1-\alpha)$-quantile of Z (obtained in tables).

- p-value of KS test: $\mathbb{P}[Z > T_n | T_n]$.

Computational issues

- In practice, how to compute T_n ?
- F^0 is non decreasing, F_n is piecewise constant, with jumps at $t_i = X_i, i = 1, \dots, n$.
- Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the reordered sample.
- The expression for T_n reduces to the following practical formula:

$$T_n = \max_{i=1,\dots,n} \left\{ \max \left(\left| \frac{i-1}{n} - F^0(X_{(i)}) \right|, \left| \frac{i}{n} - F^0(X_{(i)}) \right| \right) \right\}.$$

Pivotal distribution

- T_n is called a *pivotal statistic*: If H_0 is true, the distribution of T_n does not depend on the distribution of the X_i's and it is easy to reproduce it in simulations.
- Indeed, let $U_i = F^0(X_i), i = 1, \dots, n$ and let G_n be the empirical cdf of $U_1, \dots, U_n$.
- If H_0 is true, then $U_1, \dots, U_n \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$ and $T_n = \sup_{0 \leq x \leq 1} |G_n(x) - x|$.

Quantiles and p-values

- For some large integer M:
  - Simulate M i.i.d. copies $T_n^1, \dots, T_n^M$ of T_n;
  - Estimate the $(1-\alpha)$-quantile $q_\alpha^{(n)}$ of T_n by taking the sample $(1-\alpha)$-quantile $\hat{q}_\alpha^{(n,M)}$ of $T_n^1, \dots, T_n^M$.
- Test with approximate level α:

$$\delta_\alpha = \mathbb{1}\{T_n > \hat{q}_\alpha^{(n,M)}/\sqrt{n}\}.$$

- Approximate p-value of this test:

$$\text{p-value} \approx \frac{\#\{j = 1, \dots, M : T_n^j > T_n\}}{M}.$$

K-S table

Kolmogorov–Smirnov Tables

Critical values, $d_{alpha;(n)}$[a], of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, *n*	Level of significance, α			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893

Other goodness of fit tests

We want to measure the distance between two functions: $F_n(t)$ and $F(t)$. There are other ways, leading to other tests:

- Kolmogorov-Smirnov:

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

- Cramér-Von Mises:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dF(t)$$

- Anderson-Darling:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t)$$

Composite goodness of fit tests

What if I want to test: "Does X have Gaussian distribution?" but I don't know the parameters?
Simple idea: plug-in

$$\sup_{t \in \mathbb{R}} \left| F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t) \right|$$

where

$$\hat{\mu} = \bar{X}_n, \qquad \hat{\sigma}^2 = S_n^2$$

and $\Phi_{\hat{\mu}, \hat{\sigma}^2}(t)$ is the cdf of $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

In this case Donsker's theorem is *no longer valid*. This is a common and serious mistake!

Kolmogorov-Lilliefors test (1)

Instead, we compute the quantiles for the test statistic:

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)\right|$$

They do not depend on unknown parameters!

This is the Kolmogorov-Lilliefors test.

K-L table

Sample Size N	Level of Significance for $D = \text{Max}\ \lvert F^*(X) - S_N(X)\rvert$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231

Quantile-Quantile (QQ) plots (1)

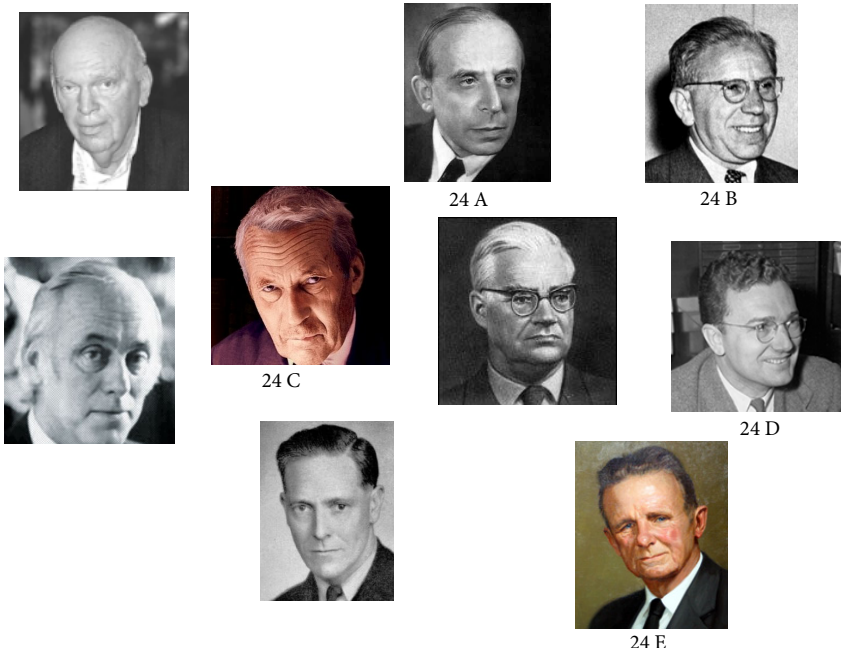
- Provide a visual way to perform GoF tests
- Not formal test but quick and easy check to see if a distribution is plausible.
- Main idea: we want to check visually if the plot of F_n is close to that of F or equivalently if the plot of F_n^{-1} is close to that of F^{-1}.
- More convenient to check if the points

$$\left(F^{-1}(\frac{1}{n}), F_n^{-1}(\frac{1}{n})\right), \left(F^{-1}(\frac{2}{n}), F_n^{-1}(\frac{2}{n})\right), \ldots, \left(F^{-1}(\frac{n-1}{n}), F_n^{-1}(\frac{n-1}{n})\right)$$

  are near the line $y = x$.
- F_n is not technically invertible but we define

$$F_n^{-1}(i/n) = X_{(i)},$$

  the ith largest observation.



Quantile-Quantile (QQ) plots (2)

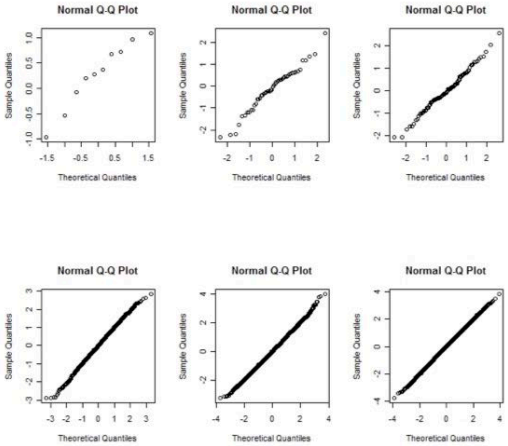


Figure 1: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

Quantile-Quantile (QQ) plots (3)

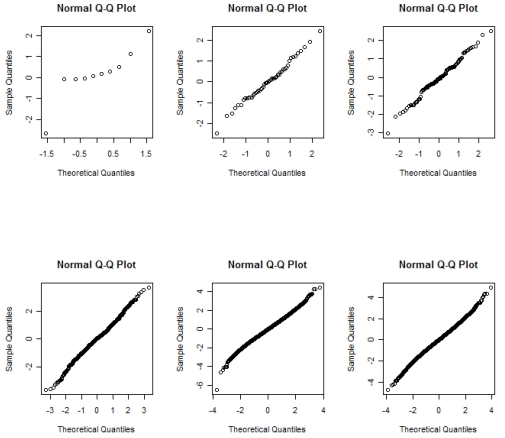


Figure 2: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a t_{15} distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

ADDED ATTRIBUTION

Slide #24A
Image from the AIP Emilio Segrè Visual Archives, Von Mises Collection. © 2019 American Institute of Physics

Slide # 24B
https://commons.wikimedia.org/wiki/File:Abraham_Wald.jpg
Citation/Attribution – Image on Wikimedia by Konrad Jacobs, Erlangen.
Copyright is MFO. (CC BY-SA) 2.0

Slide #24C
https://www.sapaviva.com/andrey-kolmogorov/
Image from Sapaviva.com
© 2017 Valentine O. Oduenyi

Slide # 24D
https://www.causeweb.org/cause/resources/fun/quotes/wellswilks-statistical-thinking
Citation/Attribution – Image © 2019 Consortium for the Advancement of Undergraduate Statistics Education. (CC BY NC SA) 4.0

Slide # 524E
https://www.statmania.info/2017/09/harald-cramer.html
Citation/Attribution – Image (c) 2016 Stat Mania

18.650 – Fundamentals of Statistics

6. Bayesian Statistics

Goals

So far, we have followed the *frequentist* approach (cf. meaning of a confidence interval).

An alternative is the **Bayesian approach**.

New concepts will come into play:

- prior and posterior distributions
- Bayes' formula
- Priors: improper, non informative
- Bayesian estimation: posterior mean, Maximum a posteriori (MAP)
- Bayesian confidence region

In a sense, Bayesian inference amounts to having a likelihood function $L_n(\theta)$ that is weighted by prior knowledge on what θ might be. This is useful in many applications.

The frequentist approach

- Assume a statistical model $(E, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$.
- We assumed that the data $X_1, \ldots, X_n$ was drawn i.i.d from $\mathbb{P}_{\theta^*}$ for some unknown **fixed** θ^*.
- When we used the MLE for example, we looked at all possible $\theta \in \Theta$.
- Before seeing the data we did not prefer a choice of $\theta \in \Theta$ over another.

The Bayesian approach

- In many practical contexts, we have a **prior belief** about θ^*
- Using the data, we want to update that belief and transform it into a **posterior belief**.

The kiss example

- Let p be the proportion of couples that turn their head to the right
- Let $X_1, \dots, X_n \overset{i.i.d}{\sim} \text{Ber}(p)$.
- In the frequentist approach, we estimated p (using the MLE), we constructed some confidence interval for p, we did hypothesis testing (e.g., $H_0 : p = .5$ v.s. $H_1 : p \neq .5$).
- Before analyzing the data, we may believe that p is likely to be close to $1/2$.
- The Bayesian approach is a tool to update our prior belief using the data.

The kiss example

- Our prior belief about p can be quantified:
- E.g., we are 90% sure that p is between $.4$ and $.6$, 95% that it is between $.3$ and $.8$, etc...
- Hence, we can model our prior belief using a distribution for p, *as if* p was random.
- In reality, the true parameter is not random ! However, the Bayesian approach is a way of modeling our belief about the parameter by doing **as if** it was random.
- E.g., $p \sim \text{Beta}(a, b)$ (*Beta distribution*). It has pdf

$$f(x) = \frac{1}{K} x^{a-1}(1-x)^{b-1} \mathbb{1}(x \in [0,1]), \quad K = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

- This distribution is called the **prior distribution**

The kiss example

- In our statistical experiment, $X_1, \dots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter p **conditionally on** p.
- After observing the available sample $X_1, \dots, X_n$, we can update our belief about p by taking its distribution conditionally on the data.
- The distribution of p conditionally on the data is called the *posterior distribution*.
- Here, the posterior distribution is

$$\text{Beta}\big(a + \sum_{i=1}^n X_i, b + n - \sum_{i=1}^n X_i\big)$$

Clinical trials

Let us revisit our clinical trial example

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

Clinical trials

- Let $\Delta_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\Delta_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.

Quantity of interest: $\theta := \Delta_d - \Delta_c$.

In practice we have a prior belief on θ. For example,

- $\theta \sim \mathrm{Unif}([100, 200])$
- $\theta \sim \mathrm{Exp}(1/100)$
- $\theta \sim \mathcal{N}(100, 300)$,
- ...

Prior and posterior

- Consider a probability distribution on a parameter space Θ with some pdf $\pi(\cdot)$: the *prior distribution*.
- Let $X_1, \ldots, X_n$ be a sample of n random variables.
- Denote by $L_n(\cdot|\theta)$ the joint pdf of $X_1, \ldots, X_n$ conditionally on θ, where $\theta \sim \pi$.
- **Remark:** $L_n(X_1, \ldots, X_n|\theta)$ is the likelihood used in the frequentist approach.
- The conditional distribution of θ given $X_1, \ldots, X_n$ is called the *posterior distribution*. Denote by $\pi(\cdot|X_1, \ldots, X_n)$ its pdf.

Bayes' formula

- Bayes' formula states that:

$$\pi(\theta|X_1, \ldots, X_n) \propto \pi(\theta) L_n(X_1, \ldots, X_n|\theta), \quad \forall \theta \in \Theta.$$

- The constant does not depend on θ:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\pi(\theta) L_n(X_1, \ldots, X_n|\theta)}{\int_\Theta L_n(X_1, \ldots, X_n|t)\pi(t)\,\mathrm{d}t}, \quad \forall \theta \in \Theta.$$

Bernoulli experiment with a Beta prior

In the Kiss example:

- $p \sim \mathrm{Beta}(a, a)$:

$$\pi(p) \propto p^{a-1}(1-p)^{a-1}, p \in (0, 1)$$

- Given p, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$, so

$$L_n(X_1, \ldots, X_n|p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

- Hence,

$$\pi(p|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}.$$

- The posterior distribution is

$$\mathrm{Beta}\left(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\right).$$

Non informative priors

- We can still use a Bayesian approach if we have no prior information about the parameter. How to pick prior π?
- Good candidate: $\pi(\theta) \propto 1$, i.e., constant pdf on Θ.
- If Θ is bounded, this is the uniform prior on Θ.
- If Θ is unbounded, this does not define a proper pdf on Θ !
- An *improper prior* on Θ is a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable.
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Examples

- If $p \sim \text{Unif}(0,1)$ and given p, $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Ber}(p)$:

$$\pi(p|X_1, \dots, X_n) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$$

  i.e., the posterior distribution is

$$\text{Beta}\left(1 + \sum_{i=1}^n X_i, 1 + n - \sum_{i=1}^n X_i\right).$$

- If $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$ and given $X_1, \dots, X_n|\theta \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$:

$$\pi(\theta|X_1, \dots, X_n) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right)$$

  i.e., the posterior distribution is

$$\mathcal{N}\left(\bar{X}_n, \frac{1}{n}\right).$$

Bayesian confidence regions

- For $\alpha \in (0,1)$, a Bayesian confidence region with level α is a random subset $\mathcal{R}$ of the parameter space Θ, which depends on the sample $X_1, \dots, X_n$, such that:

$$\mathbb{P}[\theta \in \mathcal{R}|X_1, \dots, X_n] = 1 - \alpha.$$

- Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.
- "Bayesian confidence region" and "confidence interval" are two **distinct** notions.

Bayesian estimation

- The Bayesian framework can also be used to estimate the true underlying parameter (hence, in a frequentist approach).
- In this case, the prior distribution does not reflect a prior belief: It is just an artificial tool used in order to define a new class of estimators.
- **Back to the frequentist approach:** The sample $X_1, \dots, X_n$ is associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Define a prior (that can be improper) with pdf π on the parameter space Θ.
- Compute the posterior pdf $\pi(\cdot|X_1, \dots, X_n)$ associated with π.

Bayesian estimation

- *Bayes estimator:*

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \, \mathrm{d}\pi(\theta | X_1, \ldots, X_n)$$

  This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution π (hence the superscript π).

- Another popular choice is the point that maximizes the posterior distribution, provided it is unique. It is called the MAP (maximum a posteriori):

$$\hat{\theta}^{\mathrm{MAP}} = \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta | X_1, \ldots, X_n)$$

Bayesian estimation

- In the previous examples:
  - Kiss example with prior Beta(a, a) $(a > 0)$:

$$\hat{p}^{(\pi)} = \frac{a + \sum_{i=1}^{n} X_i}{2a + n} = \frac{a/n + \bar{X}_n}{2a/n + 1}.$$

    In particular, for $a = 1/2$ (Jeffreys prior),

$$\hat{p}^{(\pi_J)} = \frac{1/(2n) + \bar{X}_n}{1/n + 1}.$$

  - Gaussian example with improper prior $\pi(\theta) \propto 1$: $\hat{\theta}^{(\pi_J)} = \bar{X}_n$.
- In each of these examples, the Bayes estimator is consistent and asymptotically normal.

- In general, the asymptotic properties of the Bayes estimator do not depend on the choice of the prior.

18.650 – Fundamentals of Statistics

7. Linear Regression

Goals

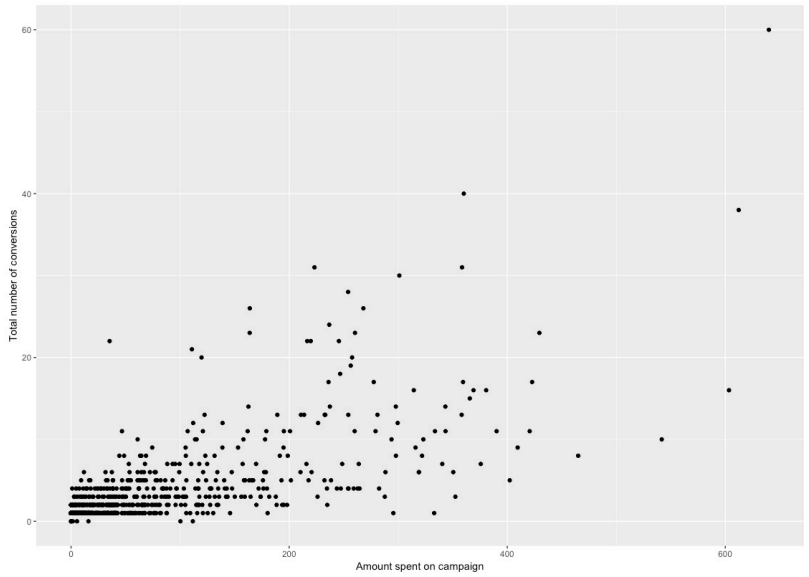
Consider two random variables X and Y. For example,

1. X is the amount of \$ spent on Facebook ads and Y is the total conversion rate
2. X is the age of the person and Y is the number of clicks

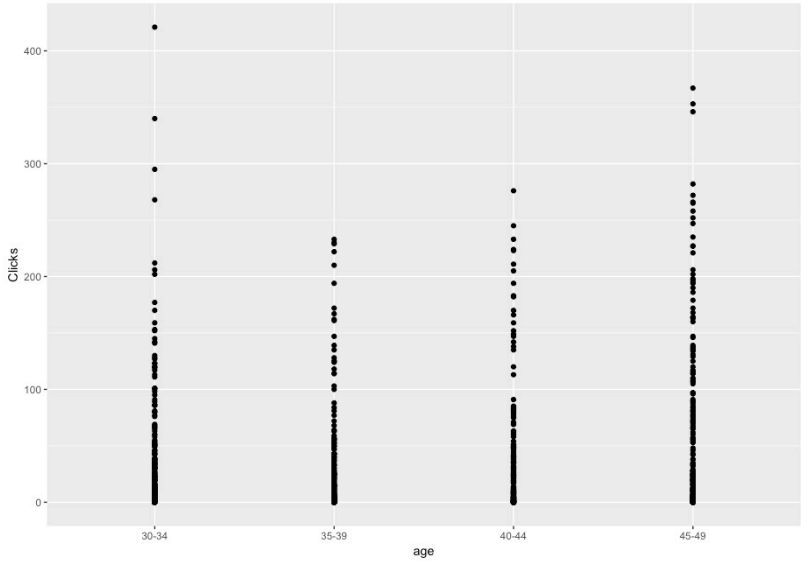
Given two random variables (X, Y), we can ask the following questions:

- How to predict Y from X?
- Error bars around this prediction?
- How much more conversions Y for an additional dollar?
- Does the number of clicks even depend on age?
- What if X is a random vector? For example, $X = (X_1, X_2)$ where X_1 is the amount of \$ spent on Facebook ads and X_2 is the duration in days of the campaign.

Conversions vs. amount spent



Clicks vs. age



Modeling assumptions

$(X_i, Y_i), i = 1, \ldots, n$ are i.i.d from some **unknown joint distribution** $\mathbb{P}$.

$\mathbb{P}$ can be described entirely by (assuming all exist)

- Either a joint PDF $h(x, y)$
- The marginal density of X $h(x) = \int h(x, y)dy$ **and** the conditional density

$$h(y|x) = \frac{h(x, y)}{h(x)}$$

$h(y|x)$ answers all our questions. It contains all the information about Y given X

Partial modeling

We can also describe the distribution only partially,e.g., using

- The expectation of Y: $\mathbb{E}[Y]$
- The conditional expectation of Y given $X = x$: $\mathbb{E}[Y|X = x]$
  The function

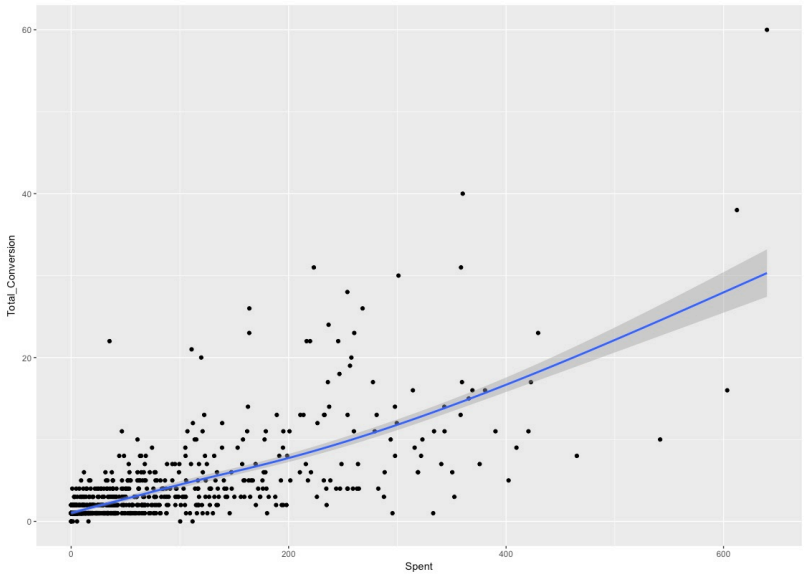
$$x \mapsto f(x) := \mathbb{E}[Y|X = x] = \int yh(y|x)dy$$

  is called **regression function**
- Other possibilities:
  - The conditional median: $m(x)$ such that

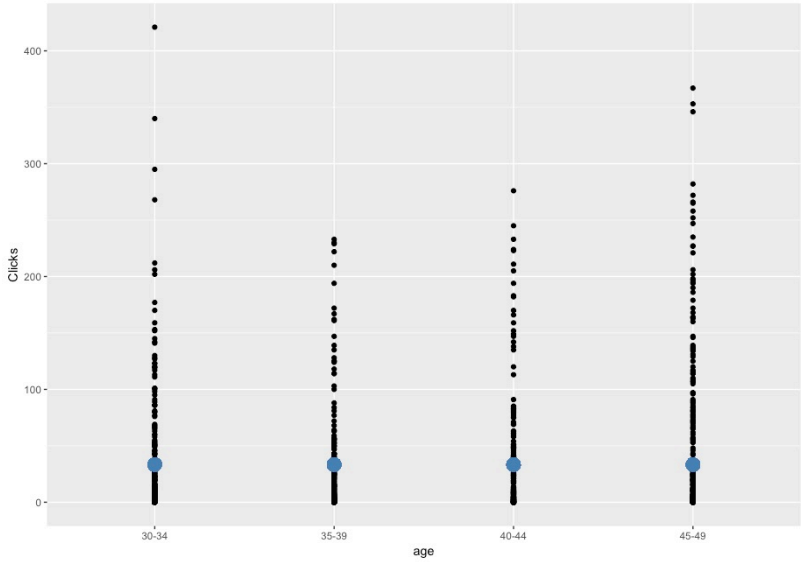
$$\int_{-\infty}^{m} h(y|x)dy = \frac{1}{2}$$

  - Conditional quantiles
  - Conditional variance (not informative about location)

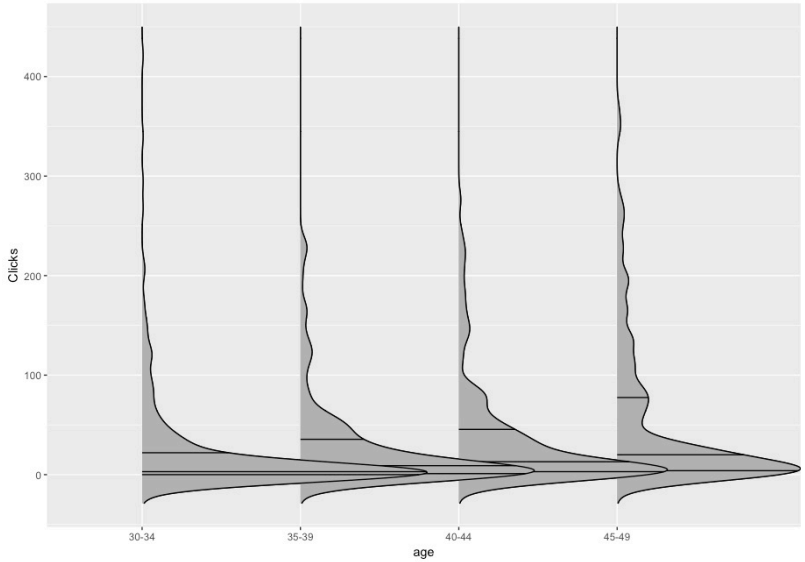
Conditional expectation and standard deviation



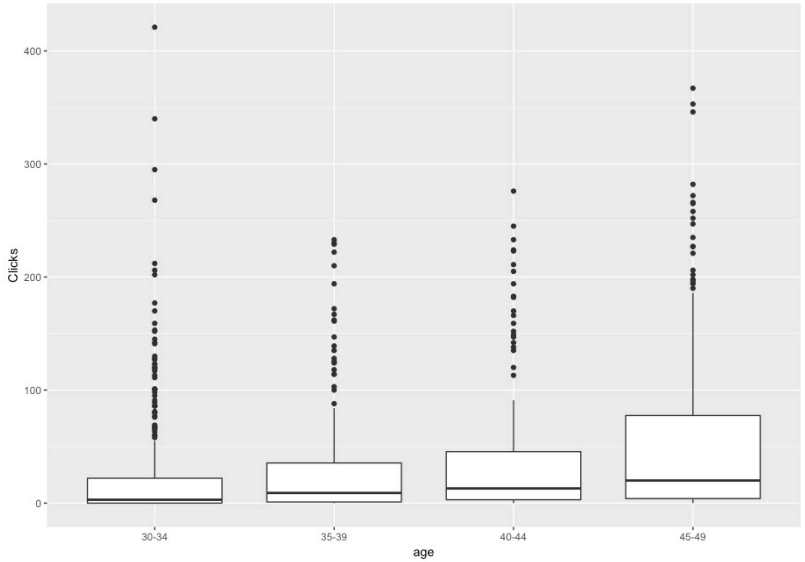
Conditional expectation



Conditional density and conditional quantiles



Conditional distribution: boxplots



Linear regression

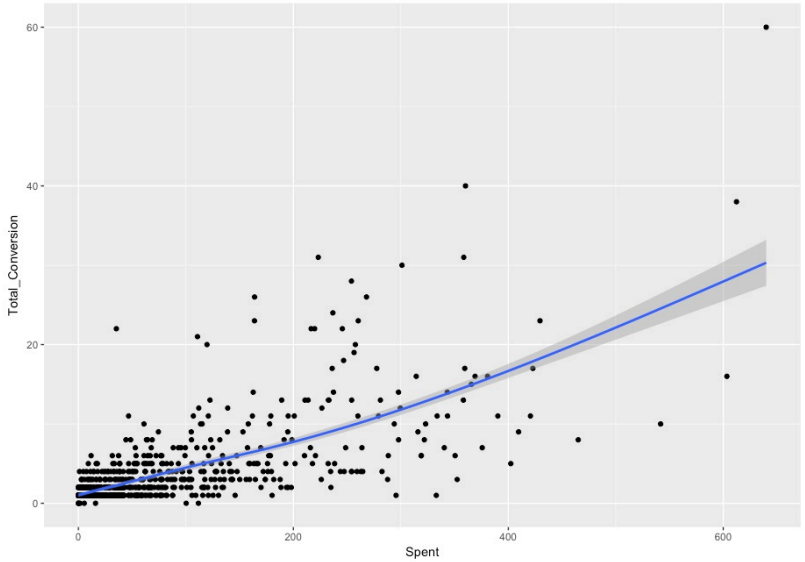
We first focus on modeling the regression function
$f(x) = \mathbb{E}[Y|X = x]$

- Too many possible regression functions f (nonparametric)
- Useful to restrict to **simple** functions that are described by a few parameters
- Simplest:

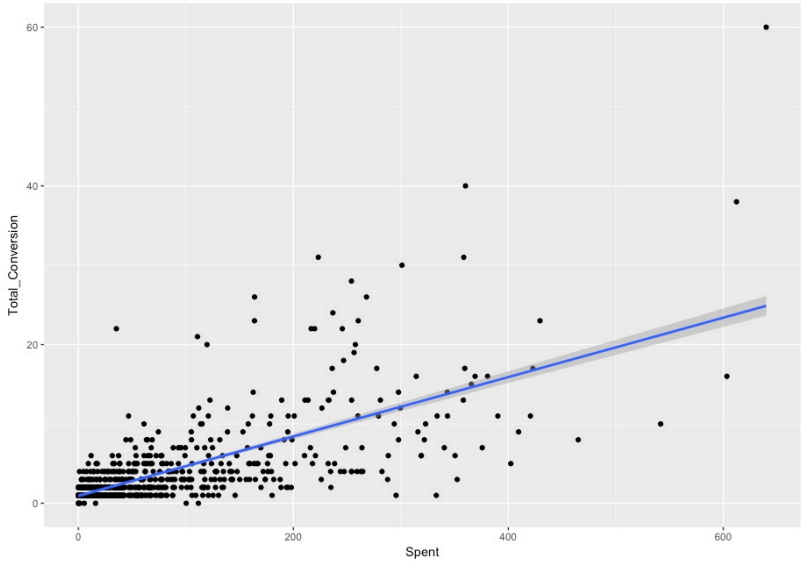
$$f(x) = a + bx \qquad \text{linear (or affine) functions}$$

Under this assumption, we talk about **linear regression**

Nonparametric regression



Linear regression



Noise

Clearly the points are not exactly on the line $x \mapsto a^* + b^* x$ if $\mathrm{var}(Y|X = x) > 0$. The random variable $\varepsilon = Y - (a^* + b^* X)$ is

called *noise* and satisfies

$$Y = a^* + b^* X + \varepsilon,$$

with

- $\mathbb{E}[\varepsilon] = 0$ and
- $\mathrm{cov}(X, \varepsilon) = 0$.

Probabilistic analysis

- Let X and Y be two real r.v. (not necessarily independent) with two moments and such that $\mathrm{var}(X) > 0$.
- The **theoretical linear regression** of Y on X is the line $x \mapsto a^* + b^* x$ where

$$(a^*, b^*) = \underset{(a,b)\in\mathbb{R}^2}{\operatorname{argmin}} \mathbb{E}\left[(Y - a - bX)^2\right]$$

- Setting partial derivatives to zero gives
  - $b^* = \dfrac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$,
  - $a^* = \mathbb{E}[Y] - b^* \mathbb{E}[X] = \mathbb{E}[Y] - \dfrac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}\ \mathbb{E}[X]$.

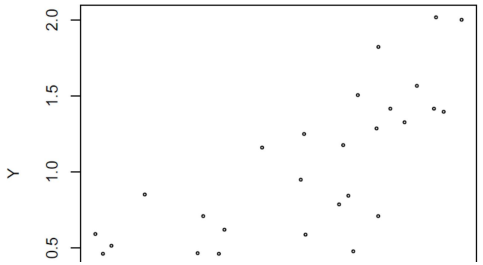
Statistical problem

In practice a^*, b^* need to be estimated from data.

- Assume that we observe n i.i.d. random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with same distribution as (X, Y):

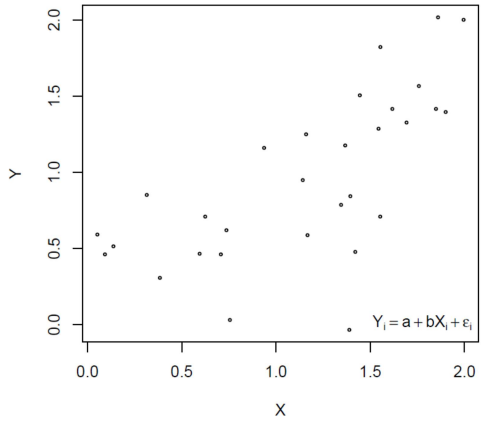
$$Y_i = a^* + b^* X_i \ + \varepsilon_i$$

- We want to estimate a^* and b^*.



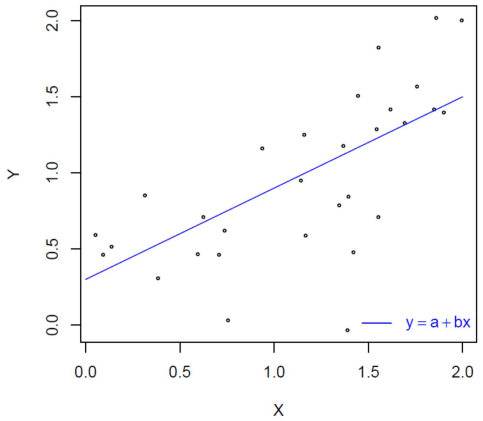
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



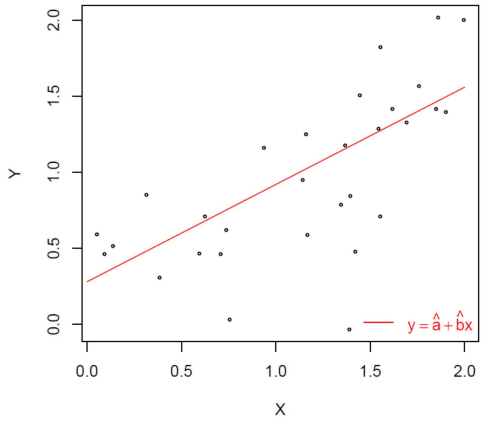
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



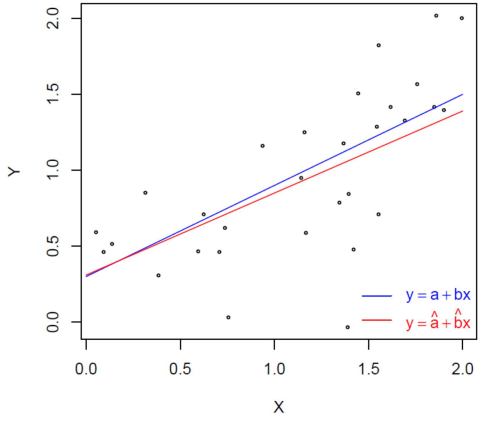
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Least squares

Definition

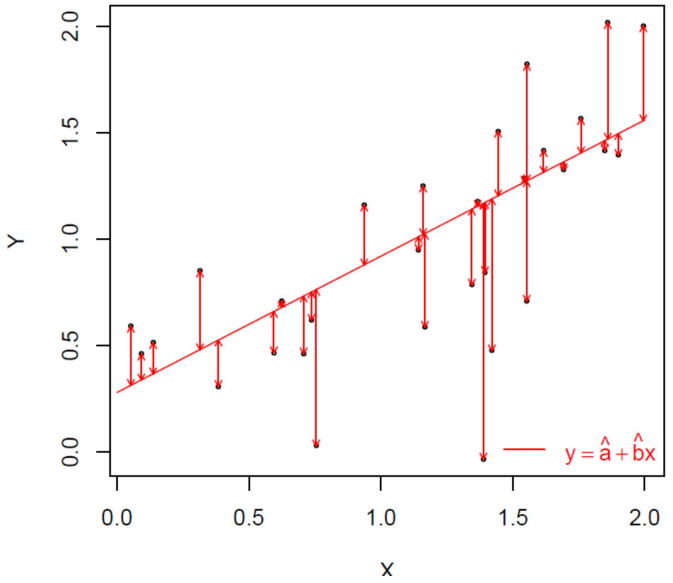
The **least squares estimator (LSE)** of (a^*, b^*) is the minimizer of the sum of squared errors:

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2.$$

$(\hat{a}, \hat{b})$ is given by

$$\hat{b} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}$$
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

Residuals



Multivariate regression

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \dots, n.$$

- Vector of **explanatory variables** or **covariates**: $\mathbf{X}_i \in \mathbb{R}^p$ (wlog, assume its first coordinate is 1).
- **Response / Dependent variable**: Y_i.
- $\boldsymbol{\beta}^* = (a^*, \mathbf{b}^{*\top})^\top$; $\beta_1^* (= a^*)$ is called the **intercept**.
- $\{\varepsilon_i\}_{i=1,\dots,n}$: noise terms satisfying $\mathrm{cov}(\mathbf{X}_i, \varepsilon_i) = \mathbf{0}$.

Definition

The **least squares estimator (LSE)** of $\boldsymbol{\beta}^*$ is the minimizer of the sum of square errors:

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2$$

LSE in matrix form

- Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$.
- Let $\mathbb{X}$ be the $n \times p$ matrix whose rows are $\mathbf{X}_1^\top, \dots, \mathbf{X}_n^\top$ ($\mathbb{X}$ is called the **design matrix**).
- Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ (unobserved noise)
- $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\beta}^*$ unknwon.
- The LSE $\hat{\boldsymbol{\beta}}$ satisfies:

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2.$$

Closed form solution

- Assume that $\text{rank}(\mathbb{X}) = p$.
- Analytic computation of the LSE:

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y}.$$

- Geometric interpretation of the LSE: $\mathbb{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbb{X}$:

$$\mathbb{X}\hat{\boldsymbol{\beta}} = P\mathbf{Y},$$

where $P = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top$.

Properties of LSE

- LSE = MLE
- Distribution of $\hat{\boldsymbol{\beta}}$: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}^*, \sigma^2(\mathbb{X}^\top \mathbb{X})^{-1}\right)$.
- Quadratic risk of $\hat{\boldsymbol{\beta}}$: $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] = \sigma^2 \text{tr}\left((\mathbb{X}^\top \mathbb{X})^{-1}\right)$.
- Prediction error: $\mathbb{E}\left[\|\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}\|_2^2\right] = \sigma^2(n-p)$.
- Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}\|_2^2$.

Theorem

- $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \hat{\sigma}^2$.

Statistical inference

To make inference (confidence regions, tests) we need more assumptions.

Assumptions:

- The design matrix $\mathbb{X}$ is deterministic and $\text{rank}(\mathbb{X}) = p$.
- The model is **homoscedastic**: $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- The noise vector $\boldsymbol{\varepsilon}$ is Gaussian:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

for some known or unknown $\sigma^2 > 0$.

Significance tests

- Test whether the j-th explanatory variable is significant in the linear regression ($1 \leq j \leq p$).
- $H_0 : \beta_j^* = 0$ v.s. $H_1 : \beta_j^* \neq 0$.
- If γ_j is the j-th diagonal coefficient of $(\mathbb{X}^\top \mathbb{X})^{-1}$ ($\gamma_j > 0$):

$$\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}.$$

- Let $T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$.
- Test with non asymptotic level $\alpha \in (0,1)$:

$$R_{j,\alpha} = \{|T_n^{(j)}| > q_{\frac{\alpha}{2}}(t_{n-p})\}$$

where $q_{\frac{\alpha}{2}}(t_{n-p})$ is the $(1-\alpha/2)$-quantile of t_{n-p}.
- We can also compute p-values.

Bonferroni's test

- Test whether a **group** of explanatory variables is significant in the linear regression.
- $H_0 : \beta_j^* = 0, \forall j \in S$ v.s. $H_1 : \exists j \in S, \beta_j^* \neq 0$, where $S \subseteq \{1, \dots, p\}$.
- *Bonferroni's test*: $R_{S,\alpha} = \bigcup_{j \in B} R_{j,\alpha/k}$, where $k = |S|$.
- This test has nonasymptotic level at most α.

Remarks

- Linear regression exhibits correlations, **NOT** causality
- Normality of the noise: One can use goodness of fit tests to test whether the residuals $\hat{\varepsilon}_i = Y_i - \mathbb{X}_i^\top \hat{\boldsymbol{\beta}}$ are Gaussian.
- Deterministic design: If $\mathbb{X}$ is not deterministic, all the above can be understood conditionally on $\mathbb{X}$, if the noise is assumed to be Gaussian, conditionally on X.

18.650 – Fundamentals of Statistics

7. Generalized linear models

Linear model

A Gaussian linear model assumes

$$Y|X = x \sim \mathcal{N}(\mu(x), \sigma^2 I),$$

And[1]

$$\mathbb{E}(Y|X = x) = \mu(x) = x^\top \beta,$$

[1] Throughout we drop the boldface notation for vectors

Components of a linear model

The two model components (that we are going to relax) are

1. Random component: the response variable Y is continuous and $Y|X = x$ is Gaussian with mean $\mu(x)$.

2. Regression function: $\mu(x) = x^\top \beta$.

Kyphosis

The Kyphosis data consist of measurements on 81 children following corrective spinal surgery. The binary response variable, Y, indicates the presence or absence of a postoperative deforming.

The three covariates are:

- $X^{(1)}$: Age of the child in month,
- $X^{(2)}$: Number of the vertebrae involved in the operation, and
- $X^{(3)}$: Start of the range of the vertebrae involved.

Write $X = (1, X^{(1)}, X^{(2)}, X^{(3)})^\top \in \mathbb{R}^4$

Kyphosis

- The response variable is binary so there is no choice: $Y|X = x$ is **Bernoulli** with expected value $\mu(x) = \mathbb{E}[Y|X = x] \in (0, 1)$
- We cannot write
$$\mu(x) = x^\top \beta$$
because the right-hand side ranges through $\mathbb{R}$.
- We need an invertible function f such that $f(x^\top \beta) \in (0, 1)$

Generalization

A generalized linear model (GLM) generalizes normal linear regression models in the following directions.

1. Random component:
$$Y|X = x \sim \text{some distribution}$$
(e.g. Bernoulli, exponential, Poisson)

2. Regression function:
$$g\big(\mu(x)\big) = x^\top \beta$$
where g called **link function** and $\mu(x) = \mathbb{E}(Y|X = x)$ is the regression function.

Predator/Prey

Consider the following model for the number of preys Y that a predator (Hawk) catches per day a predator given a number X of preys (mice) in its hunting territory.

Random component: $Y > 0$ and the variance of capture rate is known to be approximately equal to its expectation so we propose the following model:

$$Y|X = x \sim \text{Poiss}\big(\mu(x)\big)$$

Where $\mu(x) = \mathbb{E}[Y|X = x]$.

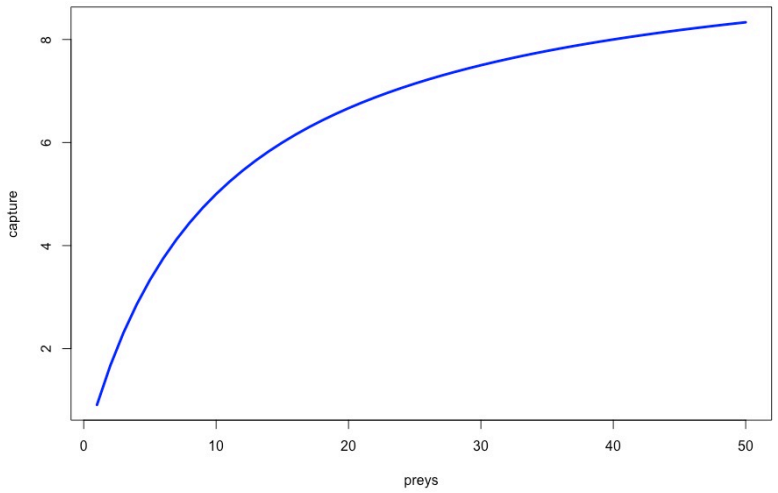
Regression function: We assume

$$\mu(x) = \frac{mx}{h + x}, \qquad \text{for some unknown } m, h > 0.$$

where:

- m is the max expected daily preys the predator can cope with
- h is the number of preys such that $\mu(h) = m/2$

The regression function $m(x)$ for $m = h = 10$



Example 2: Prey Capture Rate

Obviously $\mu(x)$ is not linear but using **reciprocal link**: $g(x) = 1/x$, the right-hand side can be made linear in the parameters:

$$g(\mu(x)) = \frac{1}{\mu(x)} = \frac{1}{\alpha} + \frac{h}{\alpha}\frac{1}{x} = \beta_0 + \beta_1 \frac{1}{x}.$$

Exponential Family

A family of distribution $\{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be a k-**parameter exponential family** on $\mathbb{R}^q$, if there exist real valued functions:

- $\eta_1, \eta_2, \cdots, \eta_k$ and B of θ,
- T_1, $T_2, \cdots, T_k$, and h of $y \in \mathbb{R}^q$ such that the density function (pmf or pdf) of $\mathbb{P}_\theta$ can be written as

$$f_\theta(y) = \exp\Big[\sum_{i=1}^{k} \eta_i(\theta) T_i(y) - B(\theta)\Big] h(y)$$

Normal distribution example

- Consider $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. The density is

  $$f_\theta(y) = \exp\Big(\frac{\mu}{\sigma^2} y - \frac{1}{2\sigma^2} y^2 - \frac{\mu^2}{2\sigma^2}\Big) \frac{1}{\sigma\sqrt{2\pi}},$$

  which forms a two-parameter exponential family with

  $$\eta_1 = \frac{\mu}{\sigma^2},\ \eta_2 = -\frac{1}{2\sigma^2},\ T_1(y) = y,\ T_2(y) = y^2,$$

  $$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}),\ h(y) = 1.$$

- When σ^2 is known, it becomes a one-parameter exponential family on $\mathbb{R}$:

  $$\eta = \frac{\mu}{\sigma^2},\ T(y) = y,\ B(\theta) = \frac{\mu^2}{2\sigma^2},\ h(y) = \frac{e^{-\frac{y^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}.$$

Examples of discrete distributions

The following distributions form **discrete** exponential families of distributions with **pmf**

- Bernoulli(p): $p^y(1-p)^{1-y}$, $y \in \{0, 1\}$
- Poisson(λ): $\frac{\lambda^y}{y!} e^{-\lambda}$, $y = 0, 1, \ldots$.

Examples of Continuous distributions

The following distributions form **continuous** exponential families of distributions with **pdf**:

- Gamma(a, b): $\frac{1}{\Gamma(a)b^a} y^{a-1} e^{-\frac{y}{b}}$;
  - above: a: shape parameter, b: scale parameter
  - reparametrize: $\mu = ab$: mean parameter

$$\frac{1}{\Gamma(a)} \left(\frac{a}{\mu}\right)^a y^{a-1} e^{-\frac{ay}{\mu}}.$$

- Inverse Gamma(α, β): $\frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\beta/y}$.
- Inverse Gaussian(μ, σ^2): $\sqrt{\frac{\sigma^2}{2\pi y^3}} e^{\frac{-\sigma^2(y-\mu)^2}{2\mu^2 y}}$.

Others: Chi-square, Beta, Binomial, Negative binomial distributions.

One-parameter canonical exponential family

- **Canonical exponential family** for $k = 1$, $y \in \mathbb{R}$

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

  for some *known* functions $b(\cdot)$ and $c(\cdot, \cdot)$.

- If ϕ is known, this is a one-parameter exponential family with θ being the canonical parameter .
- If ϕ is unknown, this may/may not be a two-parameter exponential family.
- ϕ is called **dispersion parameter**.
- In this class, we always assume that ϕ is *known*.

Normal distribution example

- Consider the following Normal density function with known variance σ^2,

$$\begin{aligned} f_\theta(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \end{aligned}$$

- Therefore $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2}$, and

$$c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right).$$

Other distributions

Table 1: Exponential Family

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of y	$(-\infty, \infty)$	$[0, -\infty)$	$\{0, 1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1 + e^\theta)$
$c(y, \phi)$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	$-\log y!$	0

Likelihood

Let $\ell(\theta) = \log f_\theta(Y)$ denote the log-likelihood function.
The mean $\mathbb{E}(Y)$ and the variance $\mathrm{var}(Y)$ can be derived from the following identities

- First identity
$$\mathbb{E}(\frac{\partial \ell}{\partial \theta}) = 0$$
- Second identity
$$\mathbb{E}(\frac{\partial^2 \ell}{\partial \theta^2}) + \mathbb{E}(\frac{\partial \ell}{\partial \theta})^2 = 0.$$

Expected value

Note that
$$\ell(\theta) = \frac{Y\theta - b(\theta)}{\phi} + c(Y;\phi),$$
Therefore
$$\frac{\partial \ell}{\partial \theta} = \frac{Y - b'(\theta)}{\phi}$$
It yields
$$0 = \mathbb{E}(\frac{\partial \ell}{\partial \theta}) = \frac{\mathbb{E}(Y) - b'(\theta)}{\phi},$$
which leads to
$$\mathbb{E}(Y) = b'(\theta).$$

Variance

On the other hand we have we have
$$\frac{\partial^2 \ell}{\partial \theta^2} + (\frac{\partial \ell}{\partial \theta})^2 = -\frac{b''(\theta)}{\phi} + \Big(\frac{Y - b'(\theta)}{\phi}\Big)^2$$
and from the previous result,
$$\frac{Y - b'(\theta)}{\phi} = \frac{Y - \mathbb{E}(Y)}{\phi}$$
Together, with the second identity, this yields
$$0 = -\frac{b''(\theta)}{\phi} + \frac{\mathrm{var}(Y)}{\phi^2},$$
which leads to
$$\mathrm{var}(Y) = b''(\theta)\phi.$$

Example: Poisson distribution

Example: Consider a Poisson likelihood,
$$f(y) = \frac{\mu^y}{y!}e^{-\mu} = \exp\big(y\log\mu - \mu - \log(y!)\big)$$
Thus,
$$\theta = \log\mu, \quad b(\theta) = \mu, \quad \phi = 1, \quad c(y,\phi) = -\log(y!),$$
So
$$\mu = e^\theta, \quad b(\theta) = e^\theta, \quad b''(\theta) = e^\theta$$

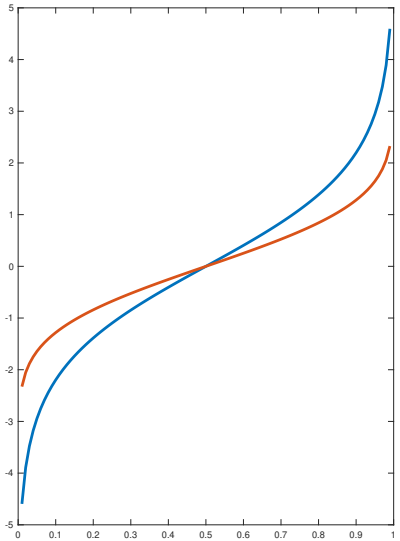
Link function

- β is the parameter of interest, and needs to appear somehow in the likelihood function to use maximum likelihood.
- A link function g relates the linear predictor $X^\top\beta$ to the mean parameter μ,
$$X^\top\beta = g(\mu).$$
- g is required to be monotone increasing and differentiable
$$\mu = g^{-1}(X^\top\beta).$$

Examples of link functions

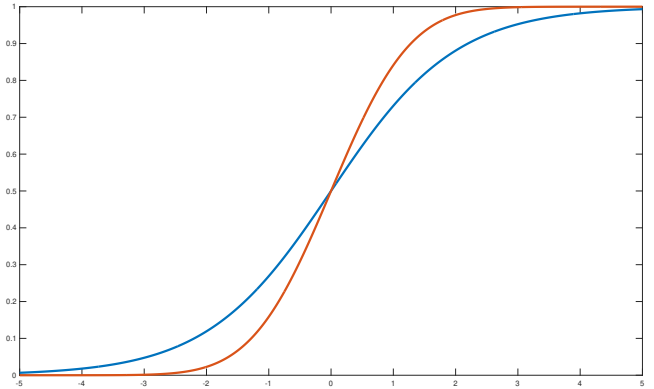
- For LM, $g(\cdot) =$ identity.
- Poisson data. Suppose $Y|X \sim \text{Poisson}(\mu(X))$.
  - $\mu(X) > 0$;
  - $\log(\mu(X)) = X^\top\beta$;
  - In general, a link function for the count data should map $(0, +\infty)$ to $\mathbb{R}$.
  - The log link is a natural one.
- Bernoulli/Binomial data.
  - $0 < \mu < 1$;
  - g should map $(0, 1)$ to $\mathbb{R}$:
  - 3 choices:
    1. logit: $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^\top\beta$;
    2. probit: $\Phi^{-1}(\mu(X)) = X^\top\beta$ where $\Phi(\cdot)$ is the normal cdf;
  - The logit link is the natural choice.

Examples of link functions for Bernoulli response



- in blue: $g_1(x) = f_1^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ (logit link)
- in red: $g_2(x) = f_2^{-1}(x) = \Phi^{-1}(x)$ (probit link)

Examples of link functions for Bernoulli response



- in blue: $f_1(x) = \frac{e^x}{1+e^x}$
- in red: $f_2(x) = \Phi(x)$ (Gaussian CDF)

Canonical Link

- The function g that links the mean μ to the canonical parameter θ is called **Canonical Link**:

$$g(\mu) = \theta$$

- Since $\mu = b'(\theta)$, the canonical link is given by

$$g(\mu) = (b')^{-1}(\mu)\,.$$

- If $\phi > 0$, the canonical link function is **strictly increasing**. `Why?`

Example: the Bernoulli distribution

- We can check that

$$b(\theta) = \log(1 + e^{\theta})$$

- Hence we solve

$$b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \mu \qquad \Leftrightarrow \qquad \theta = \log\left(\frac{\mu}{1-\mu}\right)$$

- The canonical link for the Bernoulli distribution is the **logit link**.

Other examples

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	μ
Poisson	$\exp(\theta)$	$\log \mu$
Bernoulli	$\log(1 + e^{\theta})$	$\log \frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

Model and notation

- Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ be independent random pairs such that the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family:

$$f_{\theta_i}(y_i) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\Big\}.$$

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbb{X} = (X_1, \ldots, X_n)^\top$
- Here the mean $\mu_i = \mathbb{E}[Y_i|X_i]$ is related to the canonical parameter θ_i via

$$\mu_i = b'(\theta_i)$$

- and μ_i depends linearly on the covariates through a link function g:

$$g(\mu_i) = X_i^\top \beta\,.$$

Back to β

- Given a link function g, note the following relationship between β and θ:

$$\begin{aligned} \theta_i &= (b')^{-1}(\mu_i) \\ &= (b')^{-1}(g^{-1}(X_i^\top \beta)) \equiv h(X_i^\top \beta), \end{aligned}$$

  where h is defined as

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}.$$

- Remark: if g is the **canonical** link function, h is **identity**.

Log-likelihood

- The log-likelihood is given by

$$\begin{aligned} \ell_n(\mathbf{Y}, \mathbb{X}, \beta) &= \sum_i \frac{Y_i \theta_i - b(\theta_i)}{\phi} \\ &= \sum_i \frac{Y_i h(X_i^\top \beta) - b(h(X_i^\top \beta))}{\phi} \end{aligned}$$

  up to a constant term.
- Note that when we use the **canonical** link function, we obtain the simpler expression

$$\ell_n(\mathbf{Y}, \mathbb{X}, \beta) = \sum_i \frac{Y_i X_i^\top \beta - b(X_i^\top \beta)}{\phi}$$

Strict concavity

- The log-likelihood $\ell(\theta)$ is **strictly concave** using the canonical function when $\phi > 0$. `Why?`
- As a consequence the maximum likelihood estimator is **unique**.
- On the other hand, if another parameterization is used, the likelihood function may not be strictly concave leading to **several local maxima**.

Concluding remarks

- Maximum likelihood for Bernoulli Y and the logit link is called **logistic regression**
- In general, there is no closed form for the MLE and we have to use optimization algorithms
- The asymptotic normality of the MLE also applies to GLMs.