Notes on basic probability and statistics

The sample space Ω is the set of all possible outcomes of an experiment

ex. 1: toss a coin twice $\Omega = \{HH, HT, TH, TT\}$ ← "outcomes"

ex 2: measure a distance $\Omega = [0, \infty)$

Subsets of Ω are called events

ex 1: first toss is heads: $A = \{HH, HT\}$

ex 2: distance is less than 10: $A = [0, 10)$

A function P that assigns a real number $P(A)$ to each event A is a probability measure if it satisfies: ← aka probability distribution

1. $P(A) \geq 0$ for every A
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots$ are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Two events A and B are independent if and only if

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

ex. $P(H \text{ on first toss AND } H \text{ on second toss})$

$= P(H \text{ on first toss}) \times P(H \text{ on second toss})$

If $P(B) > 0$ then the conditional probability of A given B is:

$$P(A|B) = \frac{P(AB)}{P(B)}$$ ← "A and B"

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

This is a version of Bayes Theorem

A random variable X is a mapping $X: \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

ex 1. X = number of heads

ω	$P(\{\omega\})$	$X(\omega)$
HH	1/4	2
HT	1/4	1
TH	1/4	1
TT	1/4	0

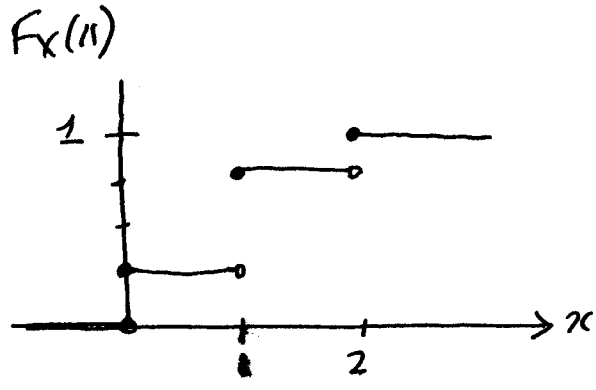
x	$P(X=x)$
0	1/4
1	1/2
2	1/4

The cumulative distribution function (CDF) is the function $F_X: \mathbb{R} \to [0,1]$ defined by:

$$F_X(x) = P(X \leq x)$$

ex 1. X = number of heads

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \le x < 1 \\ 3/4 & 1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$



X is <u>discrete</u> if it takes countably many values.

The <u>probability function</u> for X is $f_X(x) = P(X = x)$

X is <u>continuous</u> if there exists a function f_X such that $f_X(x) \ge 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ and for every $a \le b$ $P(a < X < b) = \int_a^b f_X(x)\,dx$

In this case, f_X is called a <u>probability density function</u> (PDF).

An important discrete probability function: <u>Poisson</u>

X takes values in $\{0, 1, 2, \dots\}$

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \lambda > 0 \leftarrow \text{"parameter"}$$

An important continuous probability density function: <u>Normal</u>

X takes values in $\mathbb{R}$. $N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \sigma > 0, \ \mu \in \mathbb{R}$$

The <u>expected value</u> $E(X)$ of a random variable X is:

$$E(X) = \int x \, dF_X(x) = \begin{cases} \sum_x x f_X(x) & \text{if } X \text{ is discrete} \\ \int x f_X(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

for Poisson, $E(X) = \lambda$

For Normal, $E(X) = \mu$

The <u>variance</u> $V(X)$ of a random variable is:

$$V(X) = E\left(X - E(X)\right)^2 = E(X^2) - \left(E(X)\right)^2$$

for Poisson, $V(X) = \lambda$

For Normal, $V(X) = \sigma^2$

<u>Central Limit Theorem</u>

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid), then $\bar{X}_n = \frac{1}{n} \sum X_i$ has a distribution which is approximately normal with mean μ and variance σ^2/n.

<u>Note</u> if you have f, can figure out F an vice versa

Statistical Inference

Given a sample $X_1, \dots, X_n \overset{\text{iid draws}}{\sim} F$, how do "infer" F?

(or failing that, some features of F, such as the expectation)

A parametric statistical model is a set of distributions that can be parameterized by a finite number of parameters:

Ex: a Normal model

$$\left\{ f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$$

more generally:

$$\left\{ f_X(x; \theta) : \theta \in \Theta \right\}$$

Statistical inference reduces to inferring parameters

Basic concepts

An estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n$ converges to θ as $n \to \infty$

The quality of a point estimate can be assessed by mean square error: $MSE = E(\hat{\theta}_n - \theta)^2$

$$MSE = (\bar{\theta}_n - \theta)^2 + E(\hat{\theta}_n - \bar{\theta}_n)^2$$

"bias" "variance"

EX. In a normal model, can estimate μ by:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Can show that $\hat{\mu}_n$ is consistent.

Since $\hat{\mu}_n$ is a function of $X_1, \dots, X_n$ it is itself a random variable and has a PDF. Can show that

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n)$$

This is a "sampling distribution"

The standard deviation of $\hat{\mu}_n$ (more generally $\hat{\theta}_n$) is called the standard error (SE)

$$SE = SE(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)}$$

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where a and b depend on the data such that

$$P(\theta \in C_n) \geq 1-\alpha \quad \text{for all } \theta$$

Note: C_n is random.

EX. In a normal model, a 95% confidence interval for μ is given by

($\hat{\sigma}_n$: an estimator of σ, e.g. the sample standard deviation)

$$\left(\hat{\mu}_n - 1.96\frac{\hat{\sigma}_n}{\sqrt{n}},\ \hat{\mu}_n + 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right)$$

In the normal model we know that

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

Can show that, approximately,

$$\hat{\mu}_n \sim N(\mu, \hat{\sigma}_n^2/n),$$

where $\hat{\sigma}_n$ is the sample standard deviation.

Suppose your sample comprises:

0.164, −0.051, 0.882, 0.283, −1.424

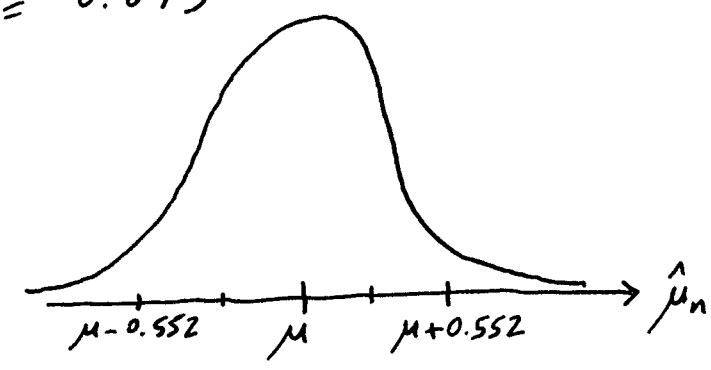
−1.120, −0.842, −1.490, −0.394, −1.700

Then $\hat{\mu}_n = \frac{\sum x_i}{10} = -0.519$

$$\hat{\sigma}_n = \sqrt{\frac{1}{9}\sum (x_i - \hat{\mu}_n)^2} = 0.873$$

$$SE = \frac{\hat{\sigma}_n}{\sqrt{n}} = 0.276$$

So $\hat{\mu}_n \sim N(\mu, 0.276^2)$



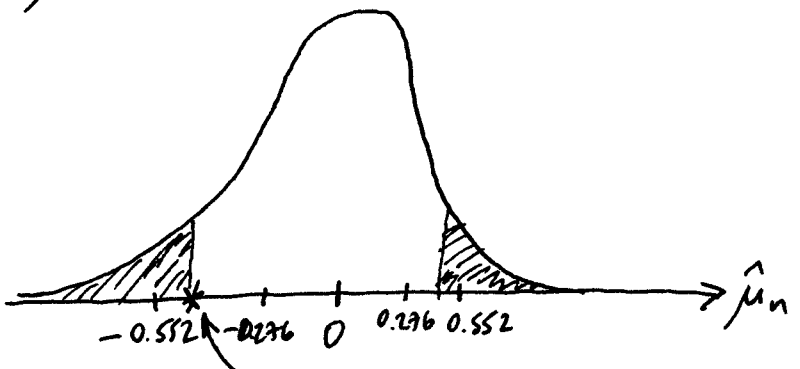
↑ the distribution of the $\hat{\mu}_n$'s that you might have gotten. Actually have one $\hat{\mu}_n$ and it is −0.519.

In <u>hypothesis testing</u> you play the following game:

Suppose μ really is 0 (or 1 or 21 or whatever)

(this is the "<u>null hypothesis</u>")

Then $\hat{\mu}_n \sim N(0, 0.276^2)$

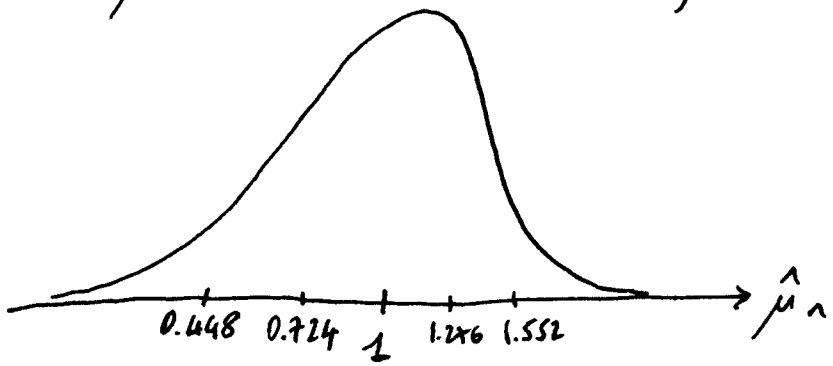


The actual observed $\hat{\mu}_n$ is here... not so close to zero!

How likely is it that you (i.e. $\hat{\mu}_n$) would be that far from the true value of μ (i.e. zero)? Answer: 0.06

This quantity is called a p-value, two-sided in this case.

How about $\mu = 1$? does that seem plausible?



The actual observed $\hat{\mu}_n$ is over here somewhere! p-value is essentially zero.

You can essentially rule out 1 as a plausible value of μ.

Hypothesis testing is a very stylized and often quite unhelpful way to think about statistical problems.

back to estimation for parametric models...

For our normal example it seems very natural to estimate μ with $\frac{\sum X_i}{n}$. (and in fact it is a fine estimator) but could use, for example, a trimmed mean or the median.

More generally, it might not be obvious how to construct an estimator. Maximum Likelihood Estimation is a general purpose scheme for producing (usually good) estimators

let $X_1, \ldots, X_n$ be iid with PDF $f_X(x;\theta)$.

The likelihood function $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Often work with the log likelihood function

$$\ell_n(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$$

The maximum likelihood estimator (MLE) denoted by $\hat{\theta}_n$ is the value of θ that maximizes $\mathcal{L}_n(\theta)$ (or equivalently maximizes $\ell_n(\theta)$)

Ex. $X_1, \dots, X_n \sim$ Bernoulli (p) (think coin tossing with a possibly biased coin)

the probability function is: $f_X(x;p) = p^x(1-p)^{1-x}$ for $x = 0, 1$.

Then $\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i;p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$

where $S = \sum_{i=1}^n X_i$

$\therefore \ell_n(p) = S \log p + (n-S)\log(1-p)$

take derivative and set equal to zero to find MLE:

$$\hat{p}_n = \frac{S}{n}$$

Ex. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$$\mathcal{L}_n(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right)$$

$$= \sigma^{-n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}\right)$$

where $\bar{X} = \frac{1}{n}\sum X_i$ and $S^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$

Take log, take derivatives with respect to μ and σ^2, and set equal to zero:

$$\hat{\mu} = \bar{X} \qquad \hat{\sigma} = S$$

Properties of MLEs

1. The MLE is consistent $\hat{\theta}_n \to$ true theta (θ^*) as $n \to \infty$

2. The MLE is asymptotically normal:

$$\frac{\hat{\theta}_n - \theta^*}{\hat{SE}} \to N(0, 1)$$

(So we know the sampling distribution and can thus get confidence intervals and p-values)

3. The MLE is the optimal estimator in a particular technical sense.

Important stuff that I am omitting for now...

- Likelihood ratio tests
- AIC / BIC etc.
- The bootstrap
- Monte Carlo
- Bayesian Methods