BASIC PROBABILITY & STATISTICS NOTES

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

  Ex 1: toss a coin twice $\Rightarrow \Omega = \{HH, HT, TH, TT\}$ ("outcomes")

  Ex 2: Reaction time to a stimulus $\Rightarrow \Omega = (0, \infty)$

- Subsets of Ω are called events

  Ex 1: First toss is heads: $A = \{HH, HT\}$

  Ex 2: Reaction time is less than 10: $A = (0, 10)$

- A function P that assigns a real number $P(A)$ to each event A in Ω is a probability (or probability distribution) if it satisfies:

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

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

- Two events A and B are independent iff (if and only if) $P(A \cap B) = P(A)\,P(B)$ ($\cap$ = and)

Ex: $P(\text{"H on first toss" and "H on second toss"}) = \frac{1}{4}$

(H = heads)

$= P(H \text{ on first toss})\,P(H \text{ on second toss}) = \frac{1}{2} \times \frac{1}{2}$

So, "H on first toss" and "H on second toss" are independent.

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

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

- Note: One can derive

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$ (a version of Bayes Thm)

- A random variable (rv), X, is a mapping (or function) $X: \Omega \to \mathbb{R}$ ← real numbers
  ↑ sample space
  that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

Ex 1: X = # of heads in two tosses of a fair coin

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

$\Rightarrow$

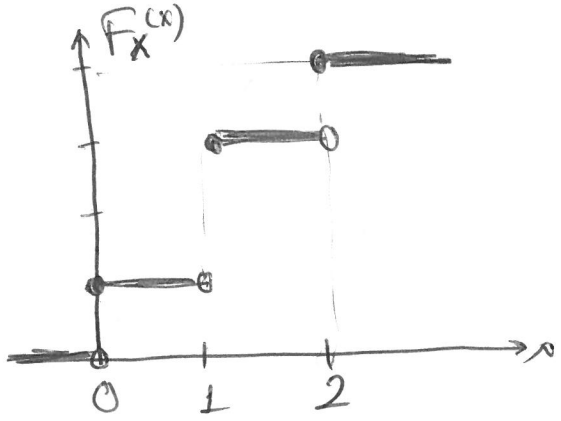
x	$P(X=x)$
0	1/4
1	1/2
2	1/4

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

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

Ex 1. X = # of heads in two tosses of the coin

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



- X is <u>discrete</u> if it takes countably many values. The <u>probability mass function</u> (pmf) for X is
$$f_X(x) = P(X = x) \quad \text{when } X \text{ is discrete}$$

- X is <u>continuous</u> if there exists a function f_X such that (s.t.)
$$f_X(x) \ge 0 \quad \text{for all } x$$
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$
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 pmf is Poisson distribution

  X takes values in $\{0,1,2,\dots\}$; $X \sim \text{Poisson}(\lambda)$

  $f_X(x) = \frac{e^{-\lambda} x^x}{x!}$, $\lambda > 0$ (λ is the parameter)

- An important continuous pdf is Normal (or Gaussian) distribution.

  X takes values in $\mathbb{R}$; $X \sim N(M, \sigma^2)$

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

- The expected (or mean) value $E(X)$ of a random variable X is

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

Ex:
For $X \sim \text{Poisson}(\lambda)$, $E(X) = \lambda$
For $X \sim N(M, \sigma^2)$, $E(X) = M$

- The variance $\text{Var}(X)$ of a rv X is

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

Ex: For $X \sim \text{Poisson}(\lambda)$, $\text{Var}(X) = \lambda$
For $X \sim N(M, \sigma^2)$, $\text{Var}(X) = \sigma^2$

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \dots, X_n$ are iid (independent identically distributed) rv's. with mean μ and variance $\sigma^2 < \infty$

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.

(i.e. $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ or

$\sqrt{n}\,\bar{X}_n \xrightarrow{d} N(\mu, \sigma^2)$ where convergence is in distribution)

- Note: If you have pmf or pdf f, you can find F and vice versa.

Statistical Inference

Given a random sample $X_1, X_2, \dots, X_n$ from F (i.e. $X_i \overset{\text{iid}}{\sim} F$ for $i = 1, 2, \dots, n$),

how to "infer" F?

(or if that is not attainable, "infer" some features (or parameters) of F, such as $E(X)$ for $X \sim F$.)

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

Ex: A normal model.

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

- More generally, a parametric model is

$$\{f_X(\cdot|\theta) \mid \theta \in \Theta\}$$

(θ: parameter (could be vector); Θ: parameter space)

Here, statistical inference boils down to inferring the parameters! (such as point estimation, interval estimation, or hypothesis testing)

Basic Concepts

Point Estimation:

- A value computed from the data, $\hat{\theta}_n$, to estimate θ is called a (point) estimator (for θ).

- An estimator $\hat{\theta}_n$ of a parameter θ is called consistent if $\hat{\theta}_n$ converges to θ (in probability) as $n \to \infty$. (That is, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$ for all $\varepsilon > 0$)

- The quality of a point estimator can be assessed, e.g., by mean square error (MSE):

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

- One can show that

$$MSE(\hat{\theta}_n) = \underbrace{(E(\hat{\theta}_n) - \theta)^2}_{\text{"bias}^2\text{"}} + \underbrace{E\left[(\hat{\theta}_n - E(\hat{\theta}_n))^2\right]}_{\text{"variance"}}$$

Ex: In a normal model, can estimate μ by:

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

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

Since $\hat{\mu}_n$ is a function of $X_1, \ldots, X_n$, it is also a rv and has a pdf (called sampling distribution)
One can show that

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

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

$$SE(\hat{\mu}_n) = \sigma/\sqrt{n}$$

In general, $SE(\hat{\theta}_n) = \sqrt{Var(\hat{\theta}_n)}$

- Interval Estimation:

A $1-\alpha$ confidence interval (CI) for a parameter θ is an interval $C_n = (L_n, U_n)$ where L_n and U_n depend on data s.t.

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

Note: C_n is random (not θ in classical statistics)

EX: In a normal model, a 95% CI for μ is given by

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

where $\hat{\sigma}_n^2$ is consistent estimator of σ^2,
e.g. the sample variance.

In the normal model, we know that

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

One can show that, approximately,

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

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

Ex: Suppose your sample consists of

0.164, −0.051, 0.882, 0.283, −1.424
−1.120, −0.342, −1.490, −0.394, −1.700

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

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

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

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

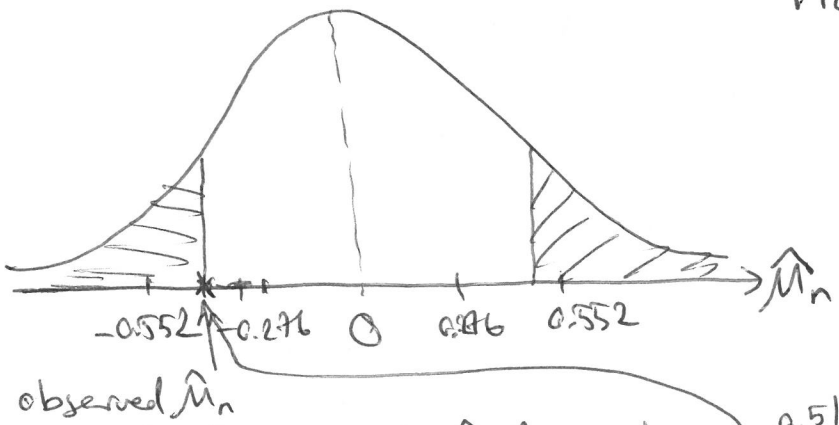
This is the distribution of the $\hat{\mu}_n$'s that you might have gotten. But from data you have one $\hat{\mu}_n$ and it is −0.519, which is a realization from $N(\mu, 0.276^2)$

Sampling distribution of $\hat{\mu}_n$ (in the $N(\mu, \sigma^2)$ model)

- Hypothesis Testing: (HT), In HT, you play the following scenario:

Suppose μ is really 0 (or 1 or 2 or whatever) (this is the "null hypothesis", $H_0: \mu = 0$)

Then $\hat{\mu}_n \overset{approx}{\sim} N(0, 0.276^2)$ ("under H_0" or "when H_0 is true")



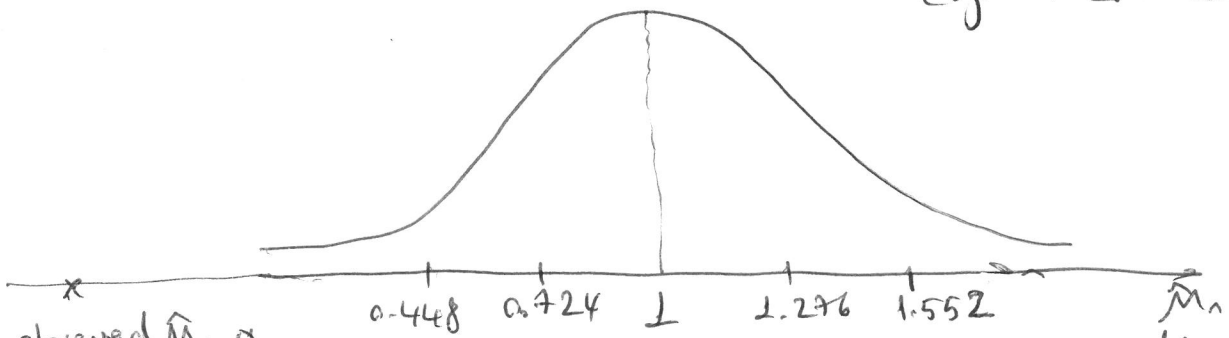
The actual "observed $\hat{\mu}_n$" is here, not so close to zero.

How likely is it that $\hat{\mu}_n$ would be that far from the true value of μ under H_0 (i.e. zero)?

Answer: 0.06

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

How about $H_0: \mu = 1$? Does this seem plausible? (given the data)



p-value $\simeq 0$. 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 a statistical problem.

Likelihood and Maximum Likelihood Estimation (MLE)

Back to estimation for parametric models...

- For our normal model it seems quite natural to estimate μ with $\hat{\mu}_n = \frac{\sum X_i}{n}$ (and in fact it is a "good" estimator), but could use, e.g., a trimmed mean or the median as an estimator for μ.

- More generally, it might not be obvious how to construct an estimator. MLE is a general purpose schema for producing (usually good) estimators.

- Let $X_1, \ldots, X_n$ be iid with pdf or pmf $f_X(x|\theta)$.

Then the joint density of $X_1, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \quad (*) \quad \text{(by iid property)}$$

Note that $(*)$ is only a function of $x_1, x_2, \ldots, x_n$ given θ and is a legitimate density function (as it is nonnegative and integrates to 1 over the joint support of X_i's.

- One can also view $(*)$ a function θ given the observed sample $x_1, \ldots, x_n$, which gives us the likelihood function $L_n(\theta)$ defined as

$$L_n(\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \qquad (**)$$

Note that $L_n(\theta)$ in (**) is a function of θ only as x_i's are observed values now and in general $L_n(\theta)$ is not a density function (as it may not integrate to 1 over Θ).
Want to find θ that maximizes $L_n(\theta)$.

- Often we work with the log likelihood function

(this is not natural logarithm, but l_n) $\rightarrow l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f_X(x_i|\theta)$

since $l_n(\theta)$ is simpler to maximize and the maximizers of $l_n(\theta)$ and $L_n(\theta)$ are same.

- The maximum likelihood estimator (MLE) denoted by $\hat{\theta}_n$ is the value of θ that maximizes $L_n(\theta)$. (or equivalently that maximizes $l_n(\theta)$)

Ex1: $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p)
(think coin tossing with a possibly biased coin).

The pmf is $f_X(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$

Then the likelihood function (for p) is

$$L_n(p) = \prod_{i=1}^{n} f_X(x_i|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

So, $l_n(p) = \log L_n(p) = \sum x_i \log p + (n - \sum x_i)\log(1-p)$

taking derivative of $l_n(p)$ with respect to p, and setting equal to zero and solving for p yields

the MLE: $\hat{p}_n = \frac{\sum x_i}{n}$

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

Then $L_n(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right)$

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

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

So, $l_n(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}$

Taking (partial) derivatives with respect to μ and σ, setting them equal to zero gives 2 equations in 2 unknowns (μ, σ). Solving for μ and σ yields the MLEs:

$$\hat{\mu}_n = \bar{X} \quad \text{and} \quad \hat{\sigma}_n = S = \sqrt{\frac{1}{n}\sum (X_i - \bar{X})^2}$$

Properties of MLE's

1. The MLE is consistent: $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$. ($\hat{\theta}_n$: MLE; θ: true theta)

2. The MLE is asymptotically normal:
$\frac{\hat{\theta}_n - \theta}{\widehat{SE}} \xrightarrow{d} N(0,1)$ (so we have the approx. sampling distr. and can get CI's and p-values)

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

Important Stuff Omitted for now:

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