Review: Probabilities

DISCRETE PROBABILITIES

Intro We have all been exposed to "informal probabilities". However it is instructive to be more precise. The goal today is to refresh our ability to reason with probabilities in simple cases. Also to explain why and when things can get complicated, and where to find the solutions in case of need.

Discrete probability space Assume we deal with a phenomenon that involves randomness, for instance it depends on k coin tosses. Let Ω be the set of all the possible sequence of coin tosses. Each $\omega \in \Omega$ is then a particular sequence of k heads or tails.

For now, we assume that Ω *contains a finite number of elements.* We make a *discrete probability space* by defining for each $\omega \in \Omega$ a *measure* $m(\omega) \in \mathbb{R}_+$. For instance this can be a count of occurences.

Since Ω is finite we can define the *elementary probabilities* $p(\omega) \triangleq \dfrac{m(\omega)}{\sum_{\omega' \in \Omega} m(\omega')}$.

Events A subset $A \subset \Omega$ is called an "event" and we define $\mathbb{P}(A) \triangleq \sum_{\omega \in A} p(\omega)$

Event language	Set language
The event A occurs	$\omega \in A$
The event A does not occur	$\omega \notin A$; $\omega \in A^c$
Both A and B occur	$\omega \in A \cap B$
A or B occur	$\omega \in A \cup B$
Either A or B occur	$\omega \in A \cup B$ and $A \cap B = \emptyset$

Essential properties $\mathbb{P}(\Omega) = 1$; $A \cap B = \emptyset \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Derived properties $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$; $\mathbb{P}(\emptyset) = 0$; $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Random variables A *random variable* X is a function of $\omega \in \Omega$ taking values from some other set $\mathcal{X}$.

That makes $\mathcal{X}$ a probability space as well: for $B \subset \mathcal{X}$, $\mathbb{P}_{\mathcal{X}}(B) = \mathbb{P}\{X \in B\} = \mathbb{P}\{\omega \,:\, X(\omega) \in B\}$.

We write X when we should write $X(\omega)$.
We write $\mathbb{P}\{X < x\}$ when we should write $\mathbb{P}\{\omega \,:\, X(\omega) < x\}$.
Same for more complicated predicates involving one or more variables.

We write $\mathbb{P}(A, B)$ instead of $\mathbb{P}(A \cap B)$, as in $\mathbb{P}\{X < x, Y = y\} = P(\{\omega : X(\omega) < x\} \cap \{\omega \,:\, Y(\omega) = y\})$.

We sometimes write $\mathbb{P}(X)$ to represent the function $x \mapsto \mathbb{P}(X = x)$.

Conditional probabilities Suppose we know event A occurs.

We can make A a probability space with the same measure: for each $\omega \in A$, $p_A(\omega) \triangleq \dfrac{m(\omega)}{\sum_{\omega' \in A} m(\omega')}$.

Then, for each $B \subset \Omega$, we define $\mathbb{P}(B|A) \triangleq \displaystyle\sum_{\omega \in B \cap A} p_A(\omega) = \frac{\sum_{\omega \in B \cap A} m(\omega)}{\sum_{\omega \in A} m(\omega)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$.

Notation: $\mathbb{P}(X|Y)$ is a function $x, y \mapsto \mathbb{P}(X = x | Y = y)$.

Bayes theorem $\mathbb{P}(B|A) \;=\; \dfrac{P(A|B)P(B)}{P(A)}$.

Chain theorem $\mathbb{P}(A_1,\ldots,A_n) \;=\; \mathbb{P}(A_1)\;\mathbb{P}(A_2|A_1)\;\mathbb{P}(A_3|A_1,A_2)\ldots\mathbb{P}(A_n|A_1\ldots A_{n-1})$

Marginalization How to compute $P(A)$ when one knows $P(A|X=x)$ for all $X \in \mathcal{X}$?

$$\mathbb{P}(A) = \sum_{x\in\mathcal{X}} \mathbb{P}(A, X=x) = \sum_{x\in\mathcal{X}} \mathbb{P}(A|X=x)\,\mathbb{P}(X=x)$$

Independent events Events A and B are independent if knowing one tells nothing about the other. That is $\mathbb{P}(A) = \mathbb{P}(A|B)$ and $\mathbb{P}(B) = \mathbb{P}(B|A)$.
Definition: A and B are independent iff $\mathbb{P}(A,B) \;=\; \mathbb{P}(A)\,\mathbb{P}(B)$.
Definition: $A_1,\ldots,A_n$ are independent iff $\forall 1 \le i_1 < \cdots < i_K \le n,\ \mathbb{P}(A_{i_1},\ldots,A_{i_K}) = \prod_{k=1}^{K} \mathbb{P}(A_{i_k})$.
This is not the same as pairwise independent: consider two coin tosses and the three events "toss 1 returns heads", "toss 2 returns heads", and "toss 1 and 2 return identical results".

Independent random variables

X and Y are independent $\iff^{\Delta}\ \forall A \subset \mathcal{X}, B \subset \mathcal{Y},\ \ \mathbb{P}(X\in A,\, Y\in B) \;=\; \mathbb{P}(X\in A)\,\mathbb{P}(Y\in B)$.

$X_1,\ldots,X_n$ are independent $\iff^{\Delta}\ \forall A_1 \subset \mathcal{X}_1,\ldots,A_n \subset \mathcal{X}_n,\ \mathbb{P}(X_1\in A_1,\ldots,X_n\in A_n) \;=\; \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$

THE MONTY-HALL PROBLEM

> You are on a game show. You are given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1. The host, who knows where the car is, opens another door, say No. 3, and reveals a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Random variables of interest The atoms are tuples (c,m,a,h):
- c $= 1,2,3$ the location of the car.
- m $= 1,2,3$ my pick of a door.
- a $=$ yes, no whether the host proposes the switch doors.
- h $= 1,2,3$ the door picked by the host.

Constructing the probability space Apply the chain rule to random variables C, M, A, H.

$$\mathbb{P}(C,M,A,H) = \mathbb{P}(C)\,\mathbb{P}(M|C)\,\mathbb{P}(A|C,M)\,\mathbb{P}(H|C,M,A)$$

$\mathbb{P}(C)$ – Equiprobability: $\mathbb{P}(C=c) = 1/3$.
$\mathbb{P}(M|C)$ – Independence: $\mathbb{P}(M|C) = \mathbb{P}(M)$ because I am not cheating.
$\mathbb{P}(A|C,M)$ – Independence: $\mathbb{P}(A|C,M) = \mathbb{P}(A|M)$ otherwise the host is cheating. Maybe he is...
$\mathbb{P}(H|C,M,A)$ – This is the complicated one.

$$\mathbb{P}(H=h|C=c,M=m,A=a) = \begin{cases} 0 & \text{if } h=c \text{ or } h=m \\ 1/2 & \text{if not zero and } m=c \\ 1 & \text{if not zero and } m\neq c \end{cases}$$

Conditionning We want $\mathbb{P}(C|M=1, A=\text{yes}, H=3, C\neq 3) = \dfrac{\mathbb{P}(C, M=1, A=\text{yes}, H=3, C\neq 3)}{\mathbb{P}(M=1, A=\text{yes}, H=3, C\neq 3)}$

We know

$$\begin{aligned}
\mathbb{P}(C=1, M=1, A=\text{yes}, H=3, C\neq 3) &= (1/3)\times P(M=1)\times P(A=\text{yes}|M=1)\times(1/2)\\
\mathbb{P}(C=2, M=1, A=\text{yes}, H=3, C\neq 3) &= (1/3)\times P(M=1)\times P(A=\text{yes}|M=1)\times 1\\
\mathbb{P}(C=3, M=1, A=\text{yes}, H=3, C\neq 3) &= 0
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{P}(C=1 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 1/3\\
\mathbb{P}(C=2 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 2/3\\
\mathbb{P}(C=3 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 0
\end{aligned}$$

Clueless host What happens if the host does not know where the car is?

$$\mathbb{P}(H=h|C=c, M=m, A=a) = \begin{cases} 0 & \text{if } h=m\\ 1/2 & \text{otherwise}\end{cases}$$

Then

$$\begin{aligned}
\mathbb{P}(C=1 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 1/2\\
\mathbb{P}(C=2 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 1/2\\
\mathbb{P}(C=3 \mid M=1, A=\text{yes}, H=3, C\neq 3) &= 0
\end{aligned}$$

Explanation 1: The host can only reveal something he knows.

Explanation 2: When $c\neq 1$, the host has now 50% chances to open a door that reveals the car. We know he did not. Therefore the conditionning operation eliminates these cases from consideration. These cases would not have been eliminated if the host had been able to choose the right door.

Cheating host What happens if $\mathbb{P}(A|C, M)$ depends on C?
– the host may decide to help us win.
– the host may decide to help us loose.

Reasoning Probabilities as a reasoning tool: the conclusions follow from the assumptions.

Observing We could also find out whether the host is clueless or cheating by observing past instances.
– If the host never reveals the car when he picks a door, he probably knows where the car is.
– If the winning frequency diverges from the theoretical probabilities, I must revise my assumptions.

Dual nature of probability theory is what makes it so interesting.

EXPECTATION AND VARIANCE

Expectation $\mathbb{E}[X] \triangleq \sum_{x\in\mathcal{X}} x\,\mathbb{P}(X=x) = \sum_{\omega\in\Omega} X(\omega)\,p(\omega)\,.$

Note: we are still in the discrete case.

Properties: $\mathbb{E}[\mathbb{1}(A)] = \mathbb{P}(A)$; $\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$; $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Conditional expectation $\mathbb{E}[X \mid Y=y] \triangleq \sum_{x\in\mathcal{X}} x\,\mathbb{P}(X=x|Y=y)$

Notations:
– $\mathbb{E}[X \mid Y=y]$ is a number.
– $\mathbb{E}[X \mid Y]$ is a random variable that depends on the realization of Y.

Properties: $\mathbb{E}[\,\mathbb{E}[X \mid Y]\,] = \mathbb{E}[X]$.

Variance $\mathrm{Var}(X) \triangleq \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \;=\; \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$

Properties: $\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$.

Standard deviation: $\mathrm{sdev}(X) = \sqrt{\mathrm{Var}(X)}$.

Covariance $\mathrm{Cov}(X,Y) \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \;=\; \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$

X and Y decorrelated $\stackrel{\triangle}{\Longleftrightarrow}$ $\mathrm{Cov}(X,Y)=0$.
X and Y decorrelated $\Longleftrightarrow$ $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$.
X and Y decorrelated $\Longleftrightarrow$ $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.
X and Y independent $\Longrightarrow$ X and Y decorrelated. The converse is not necessarily true.

Markov inequality When X is a positive real random variable, $\mathbb{P}\{X > a\} \le \dfrac{\mathbb{E}[X]}{a}$.

Proof: $\mathbb{1}\{x > a\} \le x/a$.

Chebyshev inequality $\mathbb{P}\{|X-\mathbb{E}[X]| > a\} \le \dfrac{\mathrm{Var}(X)}{a^2}$.

Proof: apply Markov to $(X-\mathbb{E}[X])^2$.

Interpretation: X cannot be too far from $\mathbb{E}[X]$.

Note: The a^{-2} tails are not very good.
– What if we apply Markov to $(X-\mathbb{E}[X])^p$ for p greater than two?
– What if we apply Markov to $e^{\alpha(X-\mathbb{E}[X])}$ for some α positive?
The latter method leads to *Chernoff bounds* with exponentially small tails.

Geometrical interpretation of covariances When the values of random variable X are vectors
$\Sigma \;\triangleq\; \mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\top\right] \;=\; \mathbb{E}\left[X\,X^\top\right] - \mathbb{E}[X]\,\mathbb{E}[X]^\top$.

This is a positive symmetric matrix. Assume it is positive definite. Applying the Markov inequality to $(X-\mathbb{E}[X])^\top \Sigma^{-1}(X-\mathbb{E}[X])$ gives $\mathbb{P}\left\{(X-\mathbb{E}[X])^\top \Sigma^{-1}(X-\mathbb{E}[X]) \ge a\right\} \le \dfrac{d}{a}$.

Since $(X-\mathbb{E}[X])^\top \Sigma^{-1}(X-\mathbb{E}[X]) = a$ is an ellipse centered on $\mathbb{E}[X]$ whose
principal axes are the eigenvector of Σ and their lengths are the square root of the eigenvalues of Σ.
The above result says that the data is very likely to lie inside the ellipse when a is large enough.
When the ellipse if very flat, that suggests linear dependencies.

Law of large numbers
Let $X_1, \dots X_n$ be independent real random variables with
identical expectations $\mathbb{E}[X_i] = \mu$ and variances $\mathrm{Var}(X_i) = \sigma^2$.

Chebyshev inequality gives $\mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{n\epsilon^2} \overset{n\to\infty}{\longrightarrow} 0.$

Note: This is the *weak law of large numbers*.

Interpretation: the average of results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

CONTINUOUS PROBABILITIES

We want to define probabilities on ensembles Ω containing an infinite number of elements. For instance, $\Omega = \mathbb{R}$. We cannot build on top of *elementary probabilities* because, in general, $p(\omega) = 0$.

Remark We cannot observe continuous probability distributions. They are abstract objects representing a limit when the number of observations grows infinitely. Consider a derivation that starts from finite observations, studies continuous distribution in the limit, and returns to finite observations. If mathematics are free of contradictions, such a derivation should be free of contradiction as well...

Bertrand Paradox Joseph Bertrand (1822-1900) is known for uncovering many probabilistic paradoxes. Here is *the paradox of the great circle.*: Consider two points drawn uniformly on the surface of a sphere. What is the probability that their angular distance is less than 10' ?

- Fix one point. Compute the proportion of the surface of the sphere that is within 10' of the first point. That is 2.1×10^{-6}.
- Fix one point. All great circles passing through the first point are equally likely to contain the second point. So fix the great circle. The probability to be within 10' of the first point on the great circle is $2 \times 10'/21600' \approx 9.2 \times 10^{-4}$.

Both methods would be perfectly correct for discrete probabilities.
The problem here is that we are conditionning on events with probability 0.
That leads to ratios of the form 0/0 which cannot be magically trusted to give the correct answer.

Kolmogorov Axioms Consider an infinite set Ω.
Assume we define a positive scalar function $\mathbb{P}$ on subsets of Ω such that:

- $\mathbb{P}(\Omega) = 1$
- $A \cap B = \emptyset \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + P(B)$

Since this is all we have used in the discrete case, we can build the same constructions.

Small problem: we need a new axiom to express that limits make sense.

- $\forall A_1 \supset A_2 \supset A_3 \supset \dots, \quad \bigcap_{i=1}^{\infty} A_i = \emptyset \implies \lim_{i\to\infty} \mathbb{P}(A_i) = 0$
- This axiom and the addition axiom can be bundled together:
  If $A_1, A_2, \dots$ are disjoint subsets, $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Big problem: No function satisfy these criteria on all subsets of Ω. See Banach-Tarski paradox.

- We need to restrict events to "nice" subsets of Ω.
  How to define "nice" could be the topic of a long lecture.
  Basically the "nice" subsets are all the intervals and all the countable unions,
  countable intersections, and differences of other "nice" subsets.

Cumulative distribution Let X be a real random variable.
The cumulative distribution $F_X(x) \triangleq \mathbb{P}\{X \leq x\}$ is an increasing function ($F_X(-\infty) = 0$, $F_X(+\infty) = 1$).

Knowing the cumulative allows us to compute the probabilities of any interval, of any countable union or intersection of intervals, and so on. Therefore it fully describes the probability distribution of X.

Let X be a vectorial random variable of dimension d.
We can similarly define a cumulative $F_X(x) \triangleq \mathbb{P}\{X_1 \leq x_1, \ldots, X_d \leq x_d\}$ with similar properties.

Density Suppose the cumulative $F_X(x)$ of the real variable X is differentiable (not always true),
then $\mathbb{P}\{A\} = \displaystyle\int_{x \in A} p(x)dx$ where the density $p(x) \triangleq F'(x)$.

Densities $p(x)$ are quite different from elementary probabilities $p(\omega)$,
even though we use the same letter for both. In particular, a density $p(x)$ can be greater than 1.

When X is a vector, we can also find densities by solving the integral equation $F(X) = \int_{X<x} p(x)dx$.
This is already more complicated. Can we generalize further?

Expectation and Integrals The discrete definition $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\,\mathbb{P}\{X = x\}$ does not work anymore.
We have to use some sort of integral. We can define expectations as $\mathbb{E}[X] = \int X\,p(x)\,dx$.
But what do we do if the density does not exist?

Breakthrough: slicing horizontally instead of vertically eliminates the need for densities.

Let X be a positive real random variable and let $\mathbb{E}[X] \triangleq \displaystyle\int X\,dP \triangleq \lim_{\epsilon \to 0} \sum_{k=0}^{\infty} \epsilon\,\mathbb{P}\{X \geq k\epsilon\}$.

For general real random variables, we can treat the positive and negative parts separately.

This defines an integral with respect to a probability distribution dP.
This is a *Lebesgue integral* and is part of the *measure theory*. In general, all the integral calculus remains the same. When there are differences, the Lebesgue integrals are often much simpler to deal with.

Notation: one often writes $\mathbb{E}[f(X)] = \int f(X)\,dP(X)$ to stress the dependency on the distribution of X.

Notation: one sometimes write $\int_A X\,dP \triangleq \int \mathbb{1}(A)\,X\,dP$.

For a positive variable X, we can write $\mathbb{E}[X] \;=\; \displaystyle\int X\,dP \;=\; \int_{x=0}^{\infty} \mathbb{P}\{X > x\}\,dx$.
The left integral is a Lebesgue integral, the right integral is a Riemann integral.

NORMAL DISTRIBUTION AND CENTRAL LIMIT THEOREM

The *normal distribution* is often called *the Gaussian distribution*.

Standard normal distribution $\mathcal{N}(0,1)$ has density $\phi(x) = \dfrac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}x^2}$.

Cumulative: $\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$ where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{t=0}^{x} e^{-t^2} dt$.

Normal distribution $X \sim \mathcal{N}(\mu, \sigma) \iff \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Density: $\frac{1}{\sigma}\,\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$.

Cumulative: $\Phi\left(\frac{x-\mu}{\sigma}\right)$

Properties: $\mathbb{E}[X] = \mu$; $\mathrm{Var}(X) = \sigma^2$; Linear combinations of normal variables are normal.

Thin tails: $\mathbb{P}\{|X-\mu| \leq \sigma\} \approx 0.67$; $\mathbb{P}\{|X-\mu| \leq 2\sigma\} \approx 0.95$; $\mathbb{P}\{|X-\mu| \leq 3\sigma\} \approx 0.997$.

Central Limit Theorem Let $X_1, \ldots X_n$ be independent real random variables with identical expectations $\mathbb{E}[X_i] = \mu$ and variances $\mathrm{Var}(X_i) = \sigma^2$.

Then the distribution of $\dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^{n} X_i$ tends to $\mathcal{N}(\mu\sqrt{n}, \sigma)$ (same mean, same sdev.)

Sums of independent real random variables tend to follow a normal distribution.
Therefore many real world quantities are normal (but not all of them!).
This is what makes the normal distribution so special.

More precisely: $\mathbb{P}\left\{\dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} \;\leq\; a\right\} \overset{n\to\infty}{\longrightarrow} \Phi(a)$.

Compare with the law of large numbers: $\mathbb{P}\left\{\left|\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}\right| \;\leq\; a\right\} \overset{n\to\infty}{\longrightarrow} 1$ for $a > 0$.

Multivariate gaussian distribution $\mathcal{N}_d(\mu, \Sigma)$ has density $\dfrac{1}{(2\pi)^{\frac{d}{2}}\,|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$

Compare with geometric intepretation of covariances.

COMPARING CLASSIFIERS

Assume we have two classifiers C_1 and C_2.
We want to use a validation set with n examples $z_1, \ldots, z_n$ to determine which one is best.

Assume the validation examples are independent and identically distributed.

Define the random variables $R_i \triangleq \begin{cases} -1 & \text{if } C_1 \text{ is correct and } C_2 \text{ is incorrect,} \\ 0 & \text{if } C_1 \text{ and } C_2 \text{ are both correct or both incorrect,} \\ +1 & \text{if } C_1 \text{ is incorrect and } C_2 \text{ is correct.} \end{cases}$

The R_i are also independent and identically distributed with expectation μ and standard deviation σ.
If $\mu > 0$ then C_2 is better than C_1. If $\mu < 0$ then C_1 is better than C_2.

In fact we have only one set of validation examples, and therefore one observation r_i for each R_i.
Let's compute $\hat{\mu} = n^{-1}\sum r_i$. Assume $\hat{\mu} > 0$. Does this mean that C_2 is better than C_1?

Method 1: Central Limit Theorem Let us assume σ is known and let $\bar{R}_n \triangleq n^{-1}\sum_{i=1}^{n} R_i$.
Because of the central limit theorem, $\dfrac{\bar{R}_n - \mu}{\sigma/\sqrt{n}}$ approximatively follows the standard normal distribution.

If μ was negative, we would have $\mathbb{P}\left\{\bar{R}_n > \hat{\mu}\right\} \;\approx\; 1 - \Phi\left(\dfrac{\hat{\mu}-\mu}{\sigma}\sqrt{n}\right) \;\leq\; 1 - \Phi\left(\dfrac{\hat{\mu}}{\sigma}\sqrt{n}\right) \;=\; \Phi\left(-\dfrac{\hat{\mu}}{\sigma}\sqrt{n}\right)$.

Translation: if C_2 was worse than C_1, the chances to observe such a positive $\hat{\mu}$ would be less than $\Phi\left(-\frac{\hat{\mu}}{\sigma}\sqrt{n}\right)$.

Method 2: Student's t-test The above method is limited because we do not know σ.
When n is very large, we can pretend that estimate $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} (r_i - \hat{\mu})^2$ is exact.
But there can be substantial differences when n is small.

William Sealy Gosset (aka Student) solved the problem. Let $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (R_i - \bar{R})^2$.
Then $\dfrac{\bar{R}_n - \mu}{S_n / \sqrt{n}}$ approximatively follows the "t-distribution with $n-1$ degrees of freedom".

So we just need to use another table.

Method 3: using Chernoff bounds The derivation is complex but the result is handy.

Let $\nu = \sum_{i=1}^{n} |r_i|$ be the number of examples for which the two classifiers disagree.

If μ was negative, we would have $\mathbb{P}\left\{\bar{R}_n > \hat{\mu}\right\} \leq e^{-\frac{n^2 \hat{\mu}^2}{\nu}}$.

Translation: if C_2 was worse than C_1, the chances to observe such a positive $\hat{\mu}$ would be less than $e^{-\frac{n^2 \hat{\mu}^2}{\nu}}$.

Alternatively: if C_2 was worse than C_1, observing $\hat{\mu} \geq \frac{1}{n} \sqrt{-\nu \log \eta}$ would have a probability $\leq \eta$.

Remark Method 3 is just for classifiers.
Methods 1 and 2 could be retargeted for other definition of the relative errors R_i.

Subtlety Compare

- "if C_2 was worse than C_1, the chances to observe such a positive $\hat{\mu}$ would be less than η."

$$\mathbb{P}\big(\bar{R}_n > \hat{\mu} \,\big|\, C_2 \text{ worse than } C_1\big) \;\leq\; \eta$$

- "The probability that C_2 is worse then C_1 given that $\bar{R}_n \geq \hat{\mu}$ is less than η."

$$\mathbb{P}\big(C_2 \text{ worse than } C_1 \,\big|\, \bar{R}_n > \hat{\mu}\big) \;\leq\; \eta$$

Can we apply the Bayes theorem? What is $\mathbb{P}\{C_2 \text{ worse than } C_1\}$ then? Fry your brain.