Probability of Replication

Replication Summer School 2026 — Day 2 of 3

Branden Fitelson

19 May 2026

Probability Calculus

Yesterday we asked what *replication* is. Today we ask: what is the *probability that a study replicates?*

This is Chapter 3 of the book [1]. To get there, we need some **probability calculus** as a tool.

Today's arc:

- **Part 1.** Probability calculus — an algebraic approach.
- **Part 2.** Applications to evidence — diagnostic testing, the base-rate fallacy, likelihood ratios, *etc.*
- **Part 3.** Miller's *probability of replication*: ARP, IRP, p_{rep}.

[Tomorrow: extension to Hierarchical Bayesian Models.]

Probability calculus is a *general theory* of probability over sentences of a logical language. It has just two ingredients:

- **Truth-functional logic** (boolean/truth-functional connectives, truth tables, possible worlds).
- **High-school algebra** (simple real arithmetic).

For n atomic sentences there are 2^n (describable) *possible worlds* w_i (*viz.*, rows of a truth-table). We assign to each possible world a *basic probability* $\Pr(w_i) = a_i$.

P	R	$\Pr(w_i)$
$\top$	$\top$	a_1
$\top$	$\bot$	a_2
$\bot$	$\top$	a_3
$\bot$	$\bot$	a_4

☞ a_i is the (rational) *degree of confidence* we assign to "w_i *is the actual world.*"

(C1) each a_i is on the unit interval,

(C2) the sum of the a_i's equals 1.

The (unconditional) probability of *any* sentence p is the sum of the a_i's over the worlds in which p is true:

$$\Pr(p) \stackrel{\text{def}}{=} \sum_{w_i \models p} \Pr(w_i)$$

$\Pr(P) = a_1 + a_2$. $\Pr(P \,\&\, R) = \Pr(P \textbf{ and } R) = a_1$.

$\Pr(P \vee R) = \Pr(P \textbf{ or } R) = a_1 + a_2 + a_3$.

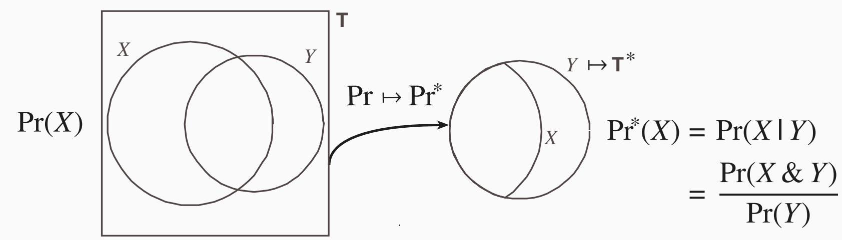
Any assignment of real numbers to the a_i which satisfies both (C1) and (C2) is called a *probability distribution*.

☞ *The algebraic approach to probability calculus*: translate Pr claims into algebraic claims involving the a_i's. Iff the resulting claim is true *for all probability distributions*, then the original claim is a *theorem of probability calculus*.

Conditional probability of X *on the supposition that* Y:

$$\Pr(X \mid Y) \stackrel{\text{def}}{=} \frac{\Pr(X \,\&\, Y)}{\Pr(Y)}, \quad \textit{if } \Pr(Y) > 0.$$

Conditioning on Y is like *zooming in on the Y-worlds, i.e.*, like *moving to a new* $\Pr^*$ *which is such that* $\Pr^*(Y) = 1$.



With our definition of CP in hand, we can express all the theorems of probability calculus. Two we'll be using:

Negation rule. $\Pr(\neg p) = 1 - \Pr(p)$.

Law of total probability. If $0 < \Pr(q) < 1$:

$$\Pr(p) = \Pr(p \mid q) \cdot \Pr(q) + \Pr(p \mid \neg q) \cdot \Pr(\neg q).$$

Recall: E is an experiment, R is its result, and P is a phenomenon (or hypothesis). If $\Pr(P), \Pr(R) > 0$, then the ratio definition immediately yields the following.

Bayes's Theorem:

$$\Pr(P \mid R) = \frac{\Pr(R \mid P) \cdot \Pr(P)}{\Pr(R)}.$$

Using the law of total probability for the denominator, we can write this in a more perspicuous and useful form:

$$\Pr(P \mid R) = \frac{\Pr(R \mid P) \cdot \Pr(P)}{\Pr(R \mid P) \cdot \Pr(P) + \Pr(R \mid \neg P) \cdot \Pr(\neg P)}.$$

Some Terminology:

- $\Pr(P)$ — *prior* (or *base rate*) of P.
- $\Pr(R \mid P)$ — *likelihood* of P on R (true positive rate).
- $\Pr(R \mid \neg P)$ — *likelihood* of $\neg P$ on R (false positive rate).
- $\Pr(P \mid R)$ — *posterior* of P on R.

☞ Bayes's Theorem is foundational in the application of probability to the evaluation of evidence.

Applications: Evidence and Diagnostic Testing

A classic application: **diagnostic testing**. Specifically, the so-called "Base-Rate Fallacy" (BRF) case:

E = a mammogram test/experiment.

R = a positive result from a mammogram test.

P = the presence of breast cancer in the patient.

$\Pr(R \mid P) = 0.85$ *(sensitivity / true positive rate)*.

$\Pr(R \mid \neg P) = 0.085$ *(false-positive rate)*.

$\Pr(P) = 0.01$ *(prior / base rate)*.

Question. A patient tests positive. What is $\Pr(P \mid R)$? (Many people — including experts — would guess *a high number*.)

We can use Bayes's theorem to calculate the posterior:

$$\Pr(P \mid R) = \frac{0.85 \cdot 0.01}{0.85 \cdot 0.01 + 0.085 \cdot 0.99} = \frac{0.0085}{0.09265} \approx 0.092.$$

Even with a positive mammogram, the patient is *still much more likely to be cancer-free* (≈ 91% *vs.* 9%).

The fact that people *underestimate* $\Pr(P \mid R)$ is called the **base-rate fallacy** (Kahneman & Tversky [5], Eddy [2]). People tend to neglect priors when the evidence is *highly relevant*.

Gigerenzer & Hoffrage [4] showed that performance on this task improves if the same problem is presented in *natural frequencies* rather than probabilities.

☞ Interestingly, these results are some of the most well replicated in modern cognitive psychology.

Algebraic Solution: solving three equations w/three unknowns determines a *unique* Pr-distribution.

$$\Pr(P) = a_1 + a_2 = 0.01, \quad \Pr(R \mid P) = \frac{a_1}{a_1 + a_2} = 0.85$$

$$\therefore a_1 = 0.0085,\ a_2 = 0.0015.$$

$$\Pr(R \mid \neg P) = \frac{a_3}{a_3 + a_4} = 0.085, \quad a_3 + a_4 = 0.99$$

$$\therefore a_3 = 0.08415,\ a_4 = 0.90585.$$

P	R	a_i
⊤	⊤	$a_1 = 0.0085$
⊤	⊥	$a_2 = 0.0015$
⊥	⊤	$a_3 = 0.08415$
⊥	⊥	$a_4 = 0.90585$

☞ $\Pr(P \mid R) = \frac{\Pr(P \,\&\, R)}{\Pr(R)} = \frac{a_1}{a_1 + a_3} = \frac{0.0085}{0.09265} \approx 0.092.$

Two Bayesian ways to model "how strongly R supports P":

Posterior. $\Pr(P \mid R)$. "How *probable* is P, *given* R?"

Likelihood ratio. "How *relevant* is R to P?"

$$LR(P,R) \overset{\text{def}}{=} \frac{\Pr(R \mid P)}{\Pr(R \mid \neg P)}.$$

"How much more probable is R if P than if $\neg P$?"

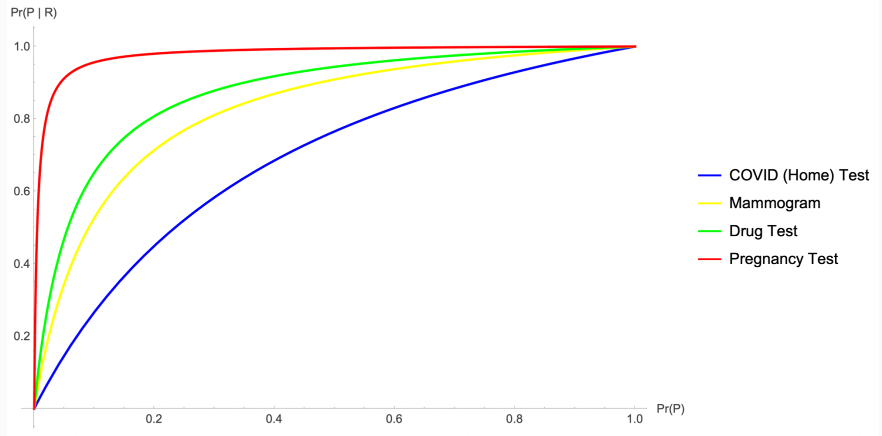
Bayes's theorem, written in *odds form*:

$$\frac{\Pr(P \mid R)}{\Pr(\neg P \mid R)} = LR(P,R) \cdot \frac{\Pr(P)}{\Pr(\neg P)}.$$

Posterior odds = likelihood ratio × prior odds.

☞ In our mammogram case, $LR(P,R) = 10$, which is pretty high. Generally, $LR(P,R) \geq 8$ is thought to indicate that R constitutes *strong evidence in favor of P* [8].

	COVID	Mammogram	Drug Test	Pregnancy Test
TPR: Pr(R | P)	0.98	0.85	0.93	0.985
FPR: Pr(R | ~P)	0.3	0.085	0.055	0.005
Relevance: LR(P, R)	3.27	10	16.9	197



Miller's Probability of Replication

Miller [6]: "What is the probability of replicating a statistically significant effect?"

Two different probabilities to distinguish:

- **Aggregate Replication Probability (ARP).** Across a research field, the probability that *a randomly chosen significant result* will replicate.
- **Individual Replication Probability (IRP).** For a *specific* observed study, the probability that *this* study will replicate.

☞ ARP is about *a field*; IRP is about *a particular study*. Both are formally defined; but, they often disagree. For simplicity, I will only discuss a *very naïve* model of ARP.

Setup. Think of E as a generic study in a field; and, let:

P: the (hypothesized) phenomenon is real

R: E produces a significant result

R': the result of a replication E' of E *is significant with the same sign*

Three parameters:

$\pi \overset{\text{def}}{=} \Pr(P)$ — proportion of real phenomena in the field.

$\alpha \overset{\text{def}}{=} \Pr(R \mid \neg P)$ — Type I error (false positive) rate.

$1 - \beta \overset{\text{def}}{=} \Pr(R \mid P)$ — power (1 − Type II error).

☞ Miller gives a Bayesian **Aggregate Replication Probability**:

$p_{\text{ARP}} \overset{\text{def}}{=} \Pr(R' \mid R)$ — given that E produces a significant result, what is the probability that E' will too?

Why ARP can be low: a (naïve) Bayesian explanation.

Among studies returning a significant result R, some are *true positives* (P) and some are *false positives* ($\neg P$).

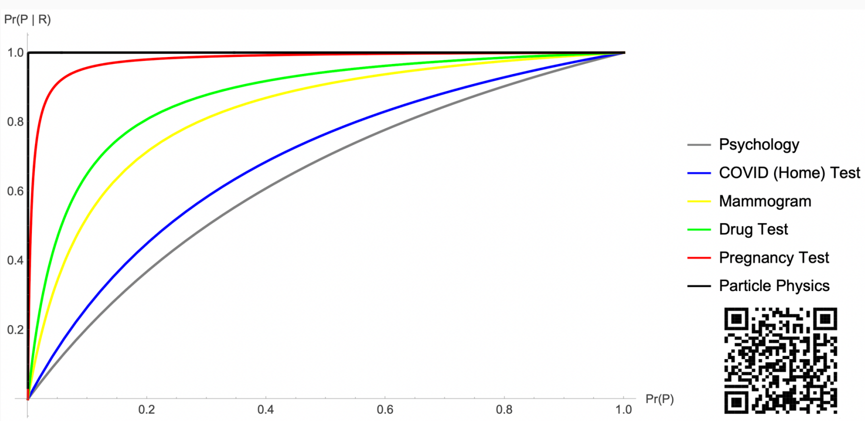
By Bayes's theorem (in our diagnostic-testing form)

$$\Pr(P \mid R) = \frac{\pi(1-\beta)}{\pi(1-\beta) + (1-\pi)\,\alpha}.$$

Suppose $\alpha = 0.05$, $1 - \beta = 0.80$, and $\pi = 0.10$. Then, we will have $\Pr(P \mid R) \approx 0.64$. About a third of the time, R will constitute *misleading evidence* in favor of P.

☞ Miller's ARP is just the BRF / diagnostic-testing problem applied to *scientific findings* — where π is the base rate of real phenomena in the field, α is the false-positive rate, and $1 - \beta$ is the sensitivity / true positive rate.

	Psychology	COVID	Mammogram	Drug Test	Pregnancy Test	Particle Physics
TPR : Pr (R | P)	0.35	0.98	0.85	0.93	0.985	0.5
FPR : Pr (R | ~P)	0.15	0.3	0.085	0.055	0.005	0.0000003
Relevance : LR (P, R)	2.33	3.27	10	16.9	197	1.67×10^6



References

[1] http://filippogambarota.github.io/replicability-book/
[2] D. Eddy, *Probabilistic reasoning in clinical medicine*, 1982.
[3] A. Gelman and J. Carlin, *Beyond power calculations: assessing type S and type M errors*, 2014.
[4] G. Gigerenzer and U. Hoffrage, *How to improve Bayesian reasoning without instruction: Frequency formats*, 1995.
[5] D. Kahneman and A. Tversky, *On the psychology of prediction*, 1973.
[6] J. Miller, *What is the probability of replicating a statistically significant effect?*, 2009.
[7] S. Pawel and L. Held, *Probabilistic forecasting of replication studies*, 2020.
[8] R. Royall, *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, 1997.