

Risk analysis

**Rough but ready tools for calculations
under variability and uncertainty**

Scott Ferson, Applied Biomathematics, scott@ramas.com

16 July 2007, International Symposium on Imprecise Probability, Prague, CZ

Risk assessments

- Environmental pollution
heavy metals, pesticides, PM_x, ozone, PCBs, EMF, RF, etc.
- Engineered systems
traffic safety, bridge design, airplanes, spacecraft, nuclear plants
- Financial investments
portfolio planning, consultation, instrument evaluation
- Occupational hazards
manufacturing and factory workers, farm workers, hospital staff
- Food safety and consumer product safety
benzene in Perrier, *E. coli* in beef, children's toys
- Ecosystems and biological resources
endangered species, fisheries and reserve management

Example: pesticides & farmworkers

- Total dose is decomposed by pathway
 - Dose from dermal exposure on hands
 - Dose from dermal exposures to rest of body
 - Dose from inhalation
 - (concentration in air, exposure duration, breathing rate, penetration factor, absorption efficiency)
- Takes account of related factors
 - Acres, gallons per acre, mixing time
 - Body mass, frequency of hand washing, etc.

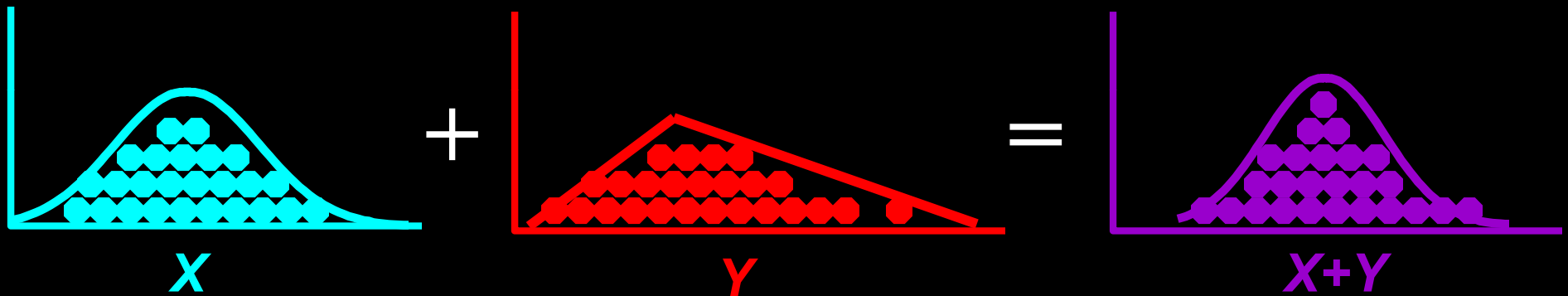
Typical problems

- Sometimes little or even no data
 - Updating is rarely used
- Very simple arithmetic (or logical) calculations
 - Occasionally, finite element meshes or differential equations
- Usually small in number of inputs
 - Nuclear power plants are the exception
- Results are important and often high profile
 - But the approach is being used ever more widely

Worst-case analysis

- Uses deterministic point values
- Mostly (but usually not all) extremes
- Actually an interval analysis
- Says how bad it could be, but not how likely that outcome is

Monte Carlo simulation



- Requires specifying the full joint distribution, i.e., all the marginals and all their dependencies
- Often we need to guess about a lot of it

Probability vs. intervals

- **Probability theory**

- Can handle likelihoods and dependence well
- Has an inadequate model of ignorance
- Lying: saying more than you really know

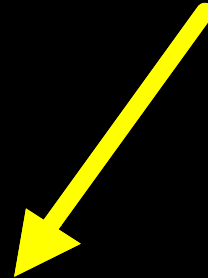
- **Interval analysis**

- Can handle epistemic uncertainty (ignorance) well
- Inadequately models frequency and dependence
- Cowardice: saying *less* than you know

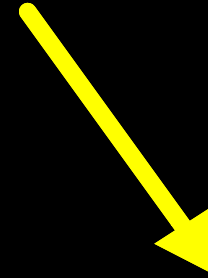
What's needed

- Reliable, conservative assessments of tail risks
- Using available information but without forcing analysts to make unjustified assumptions
- Neither computationally expensive nor intellectually taxing

Deterministic
calculation



Probabilistic
convolution



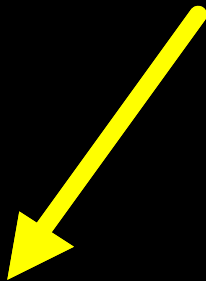
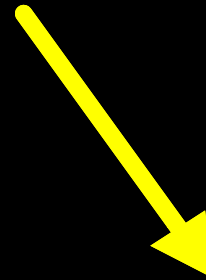
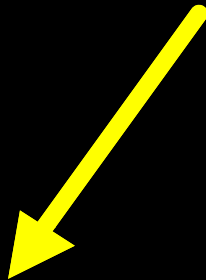
Interval
analysis

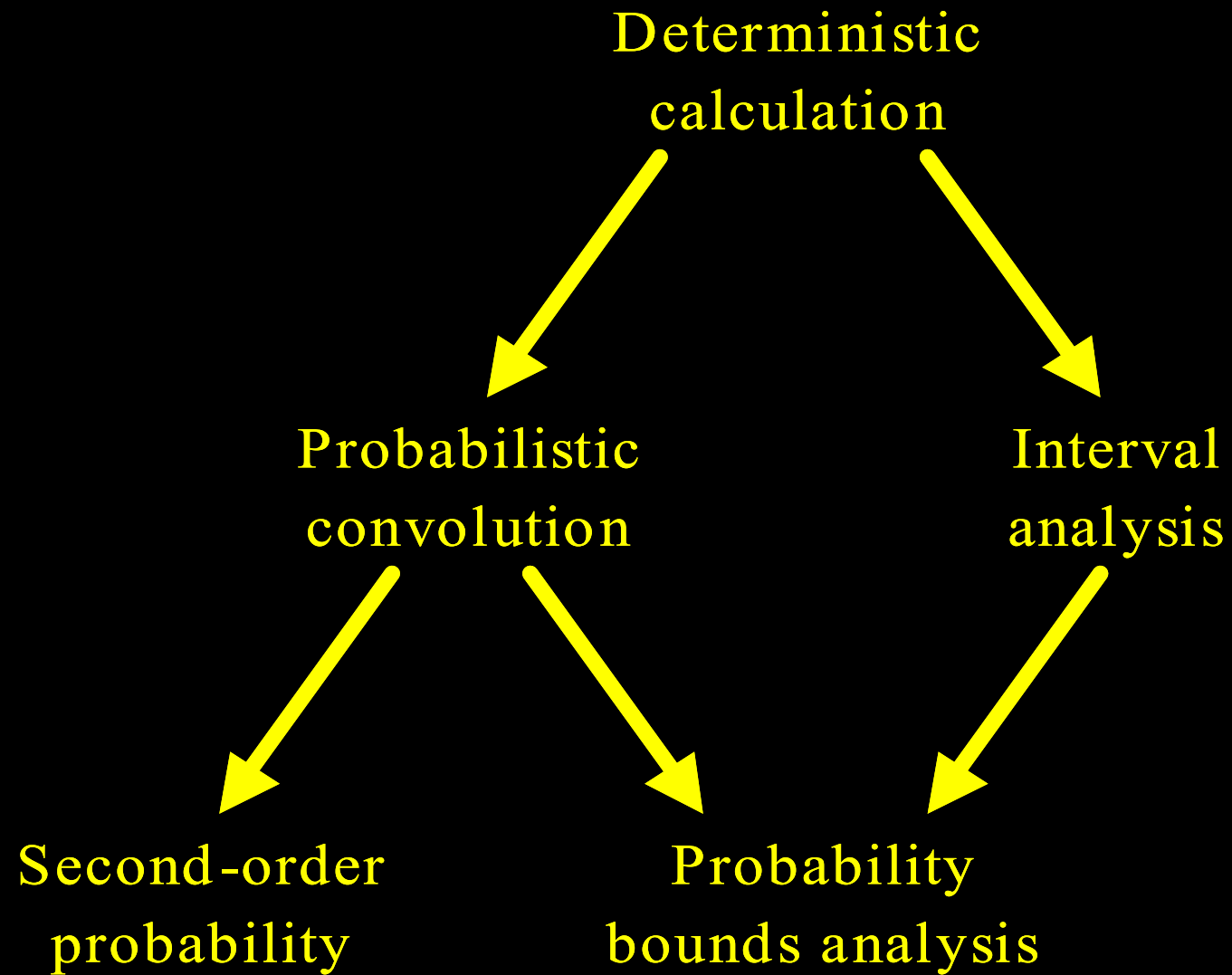
Deterministic
calculation

Probabilistic
convolution

Interval
analysis

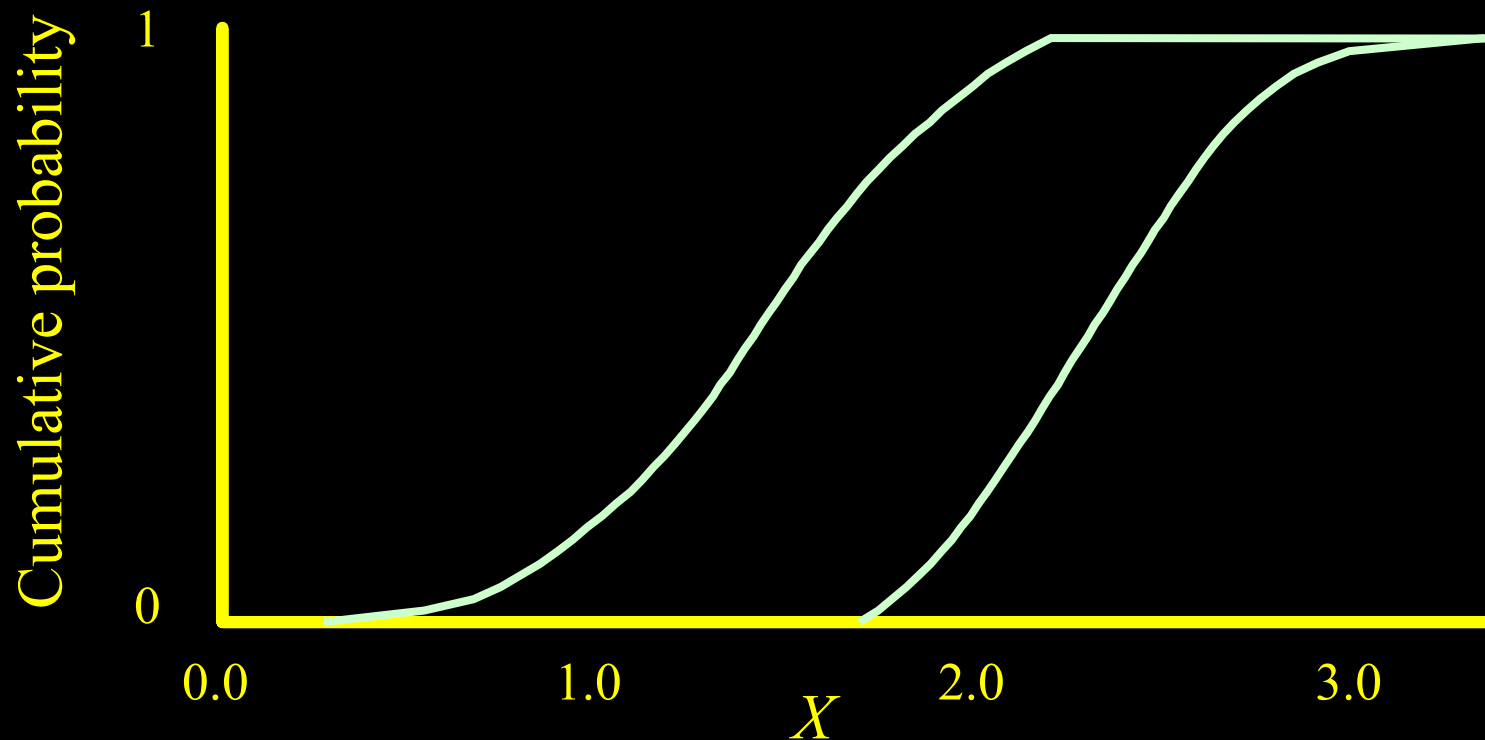
Second-order
probability





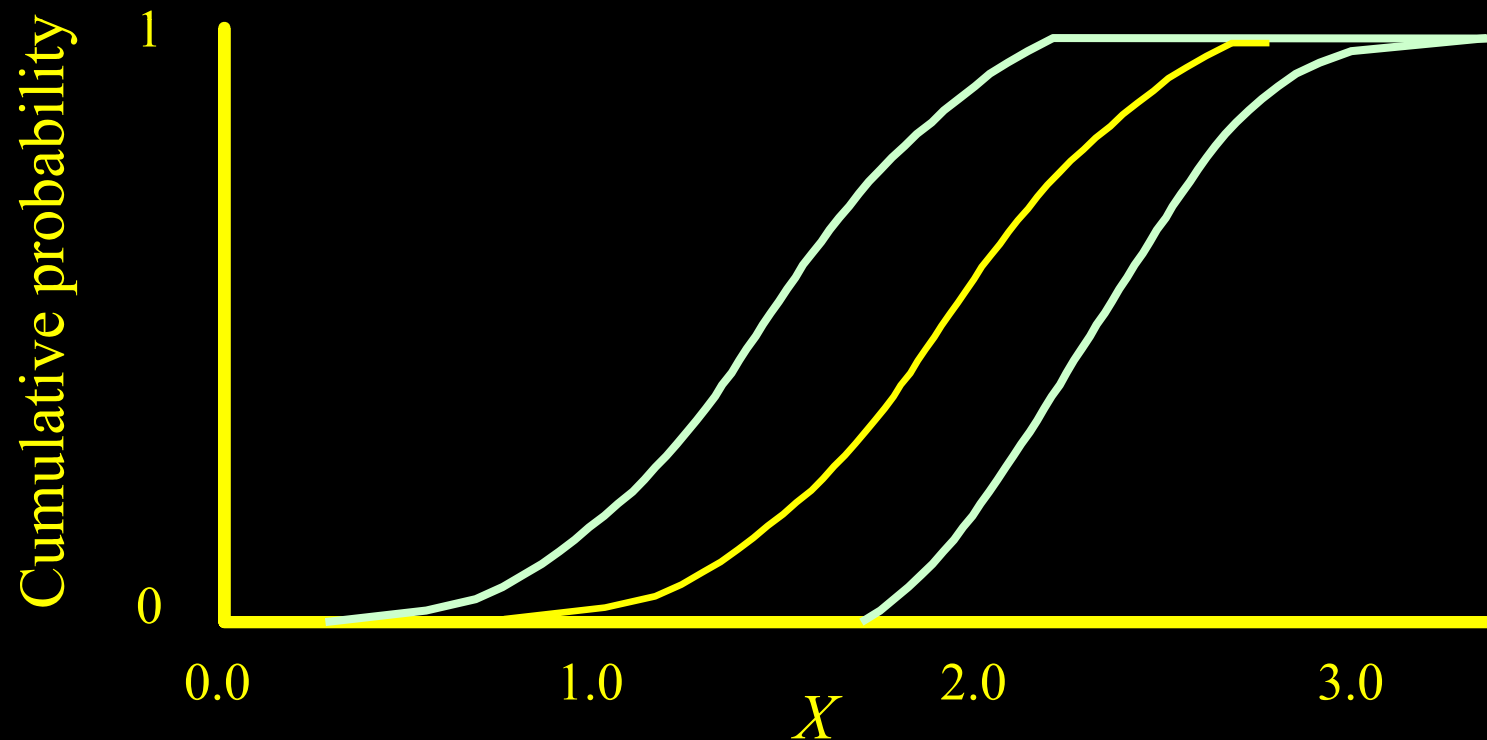
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



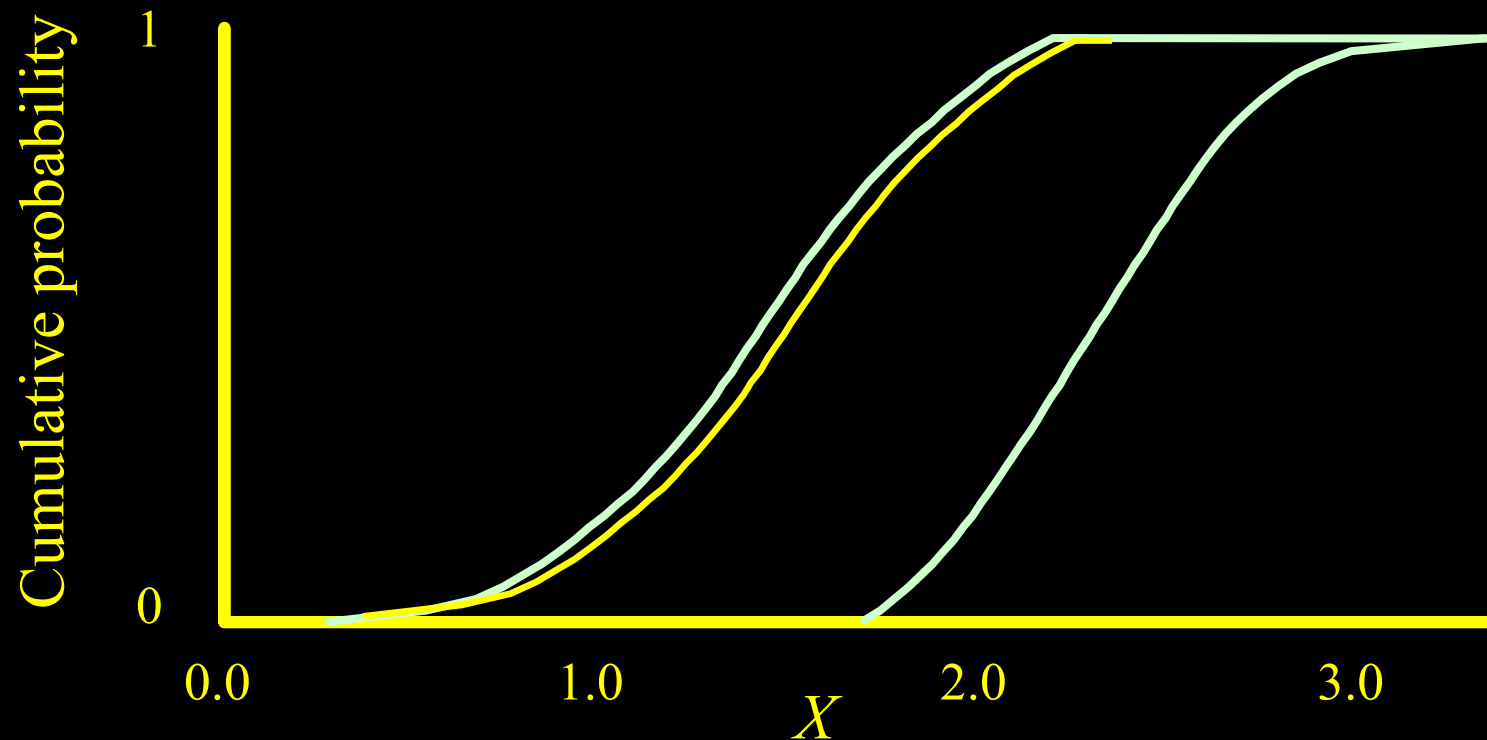
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



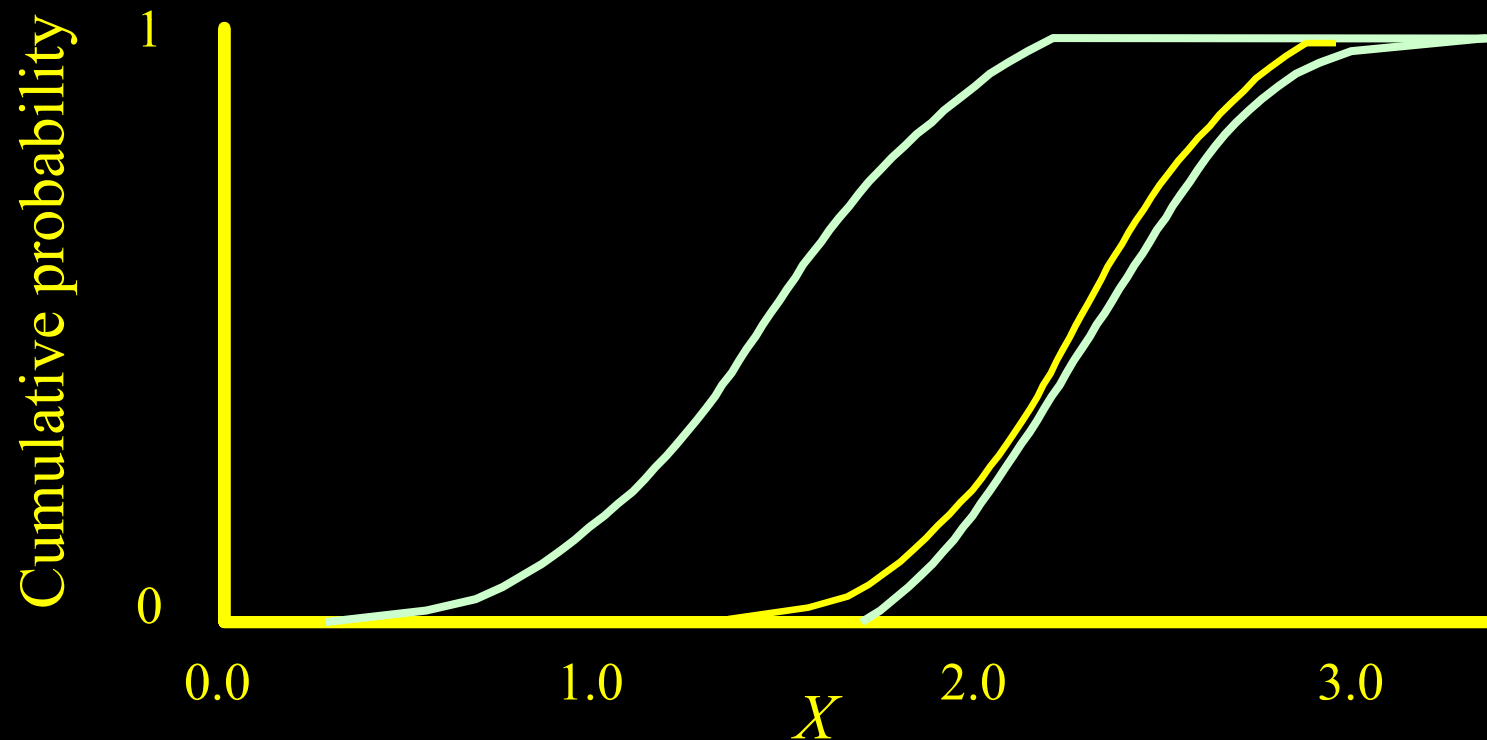
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



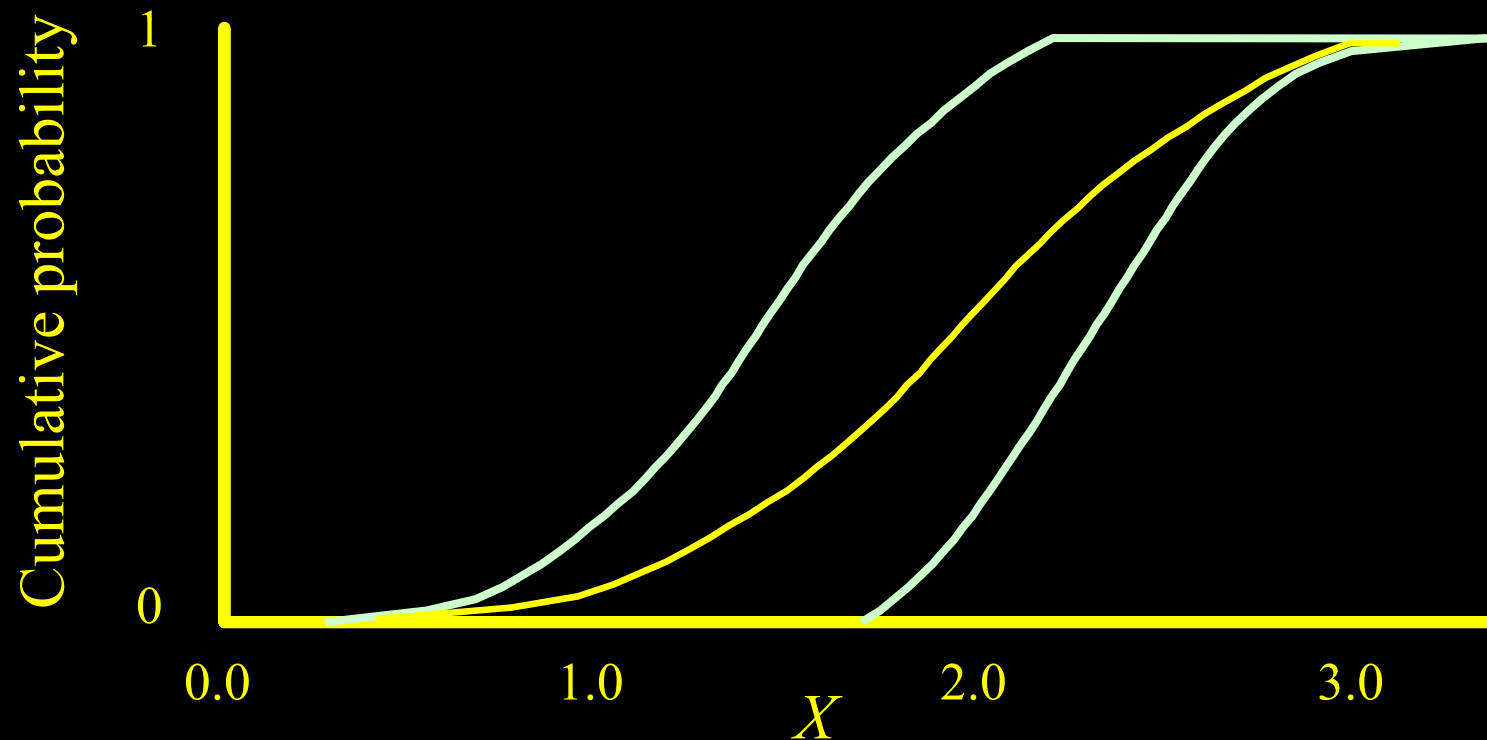
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



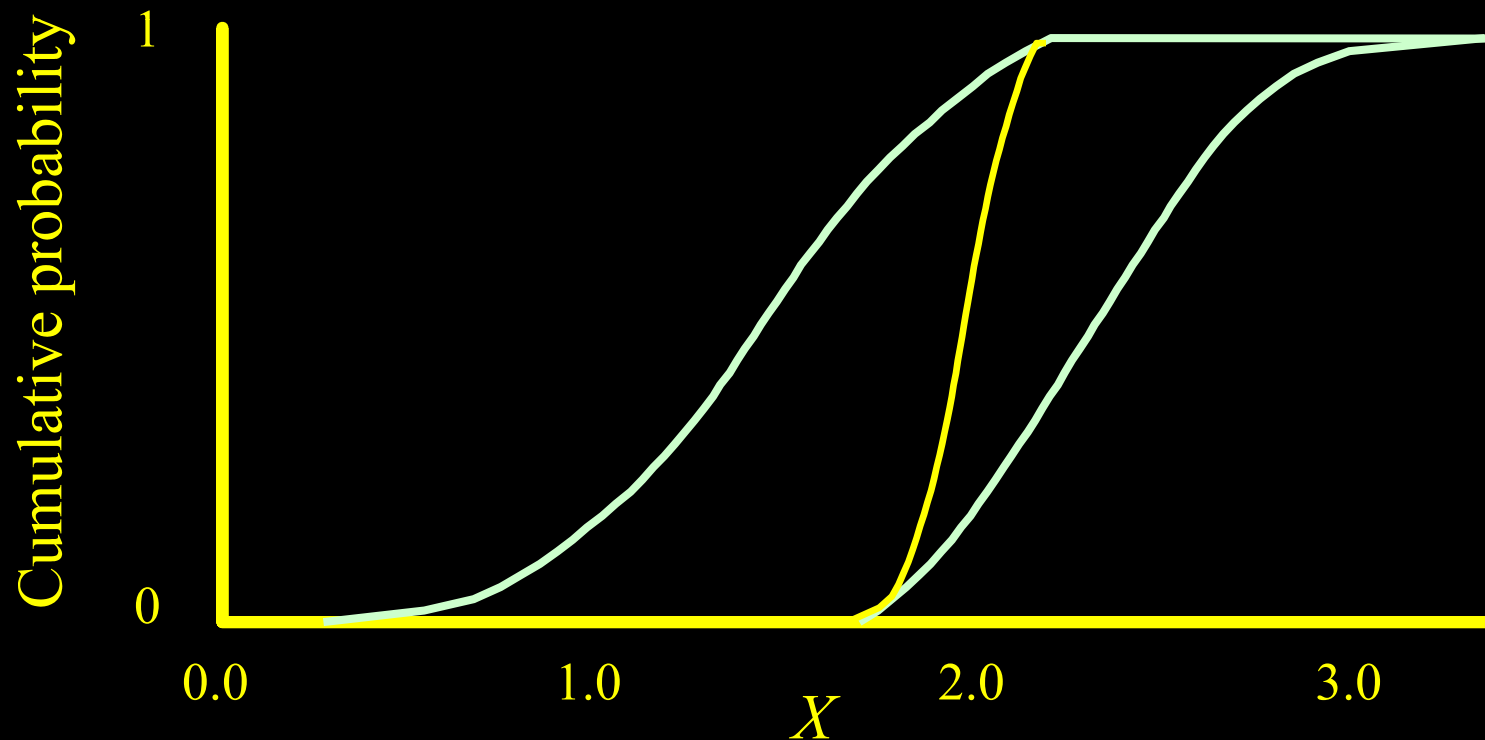
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



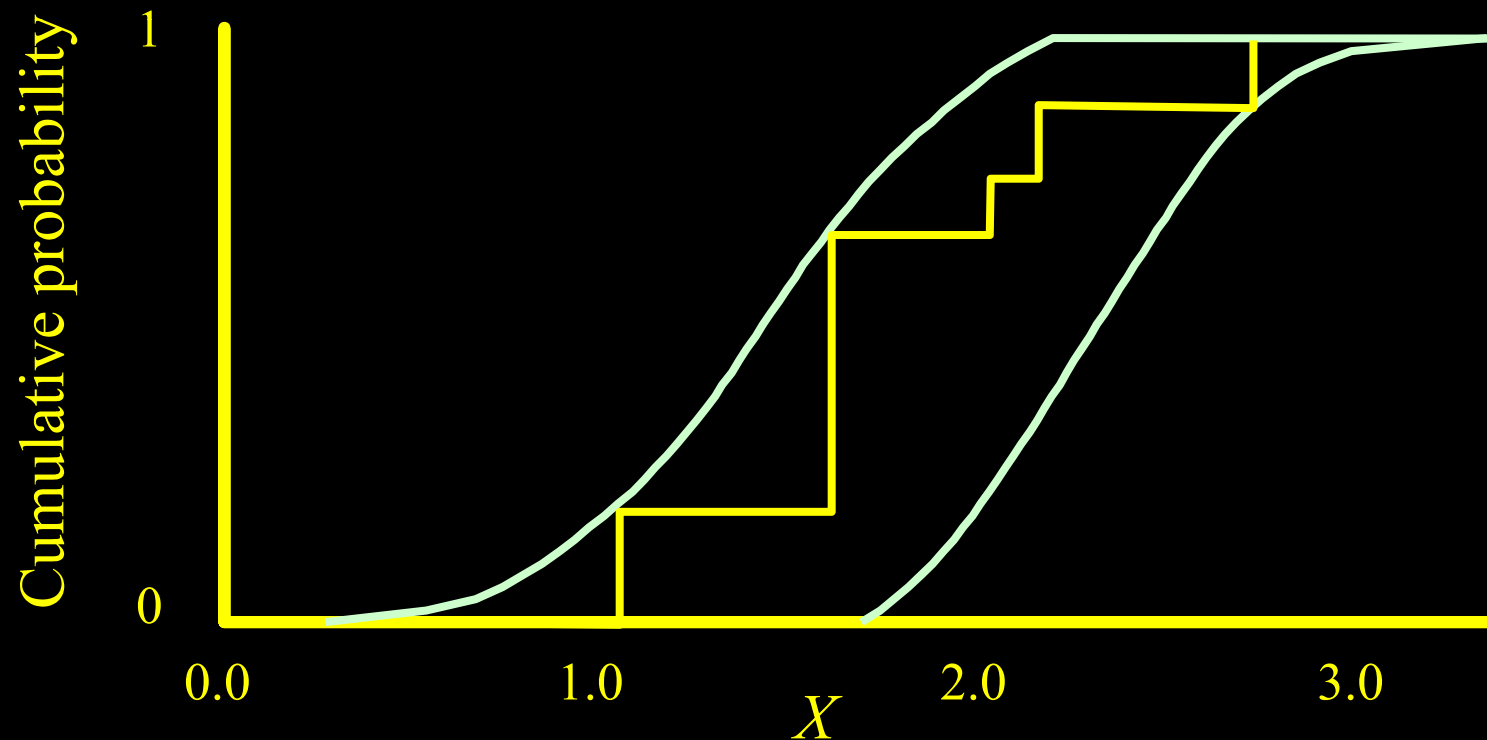
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



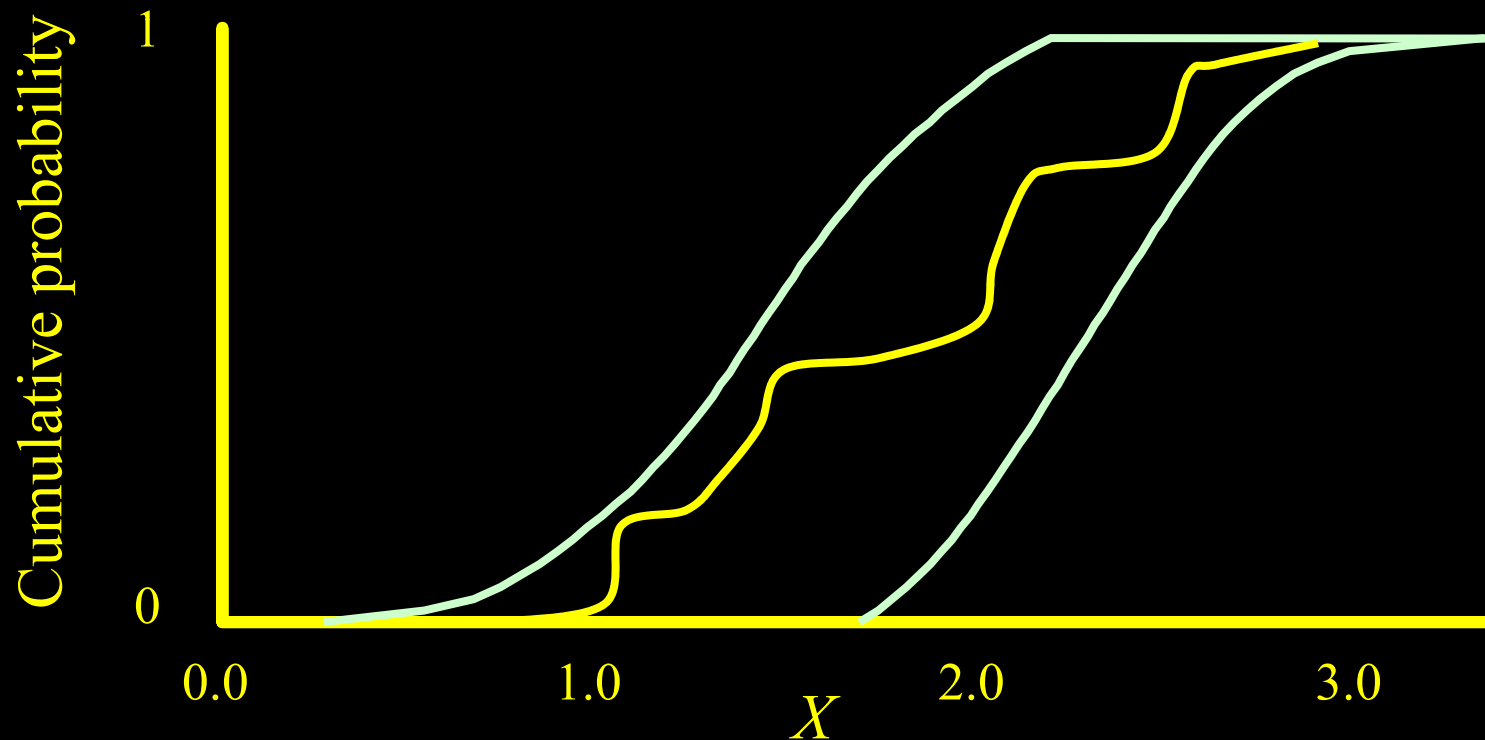
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



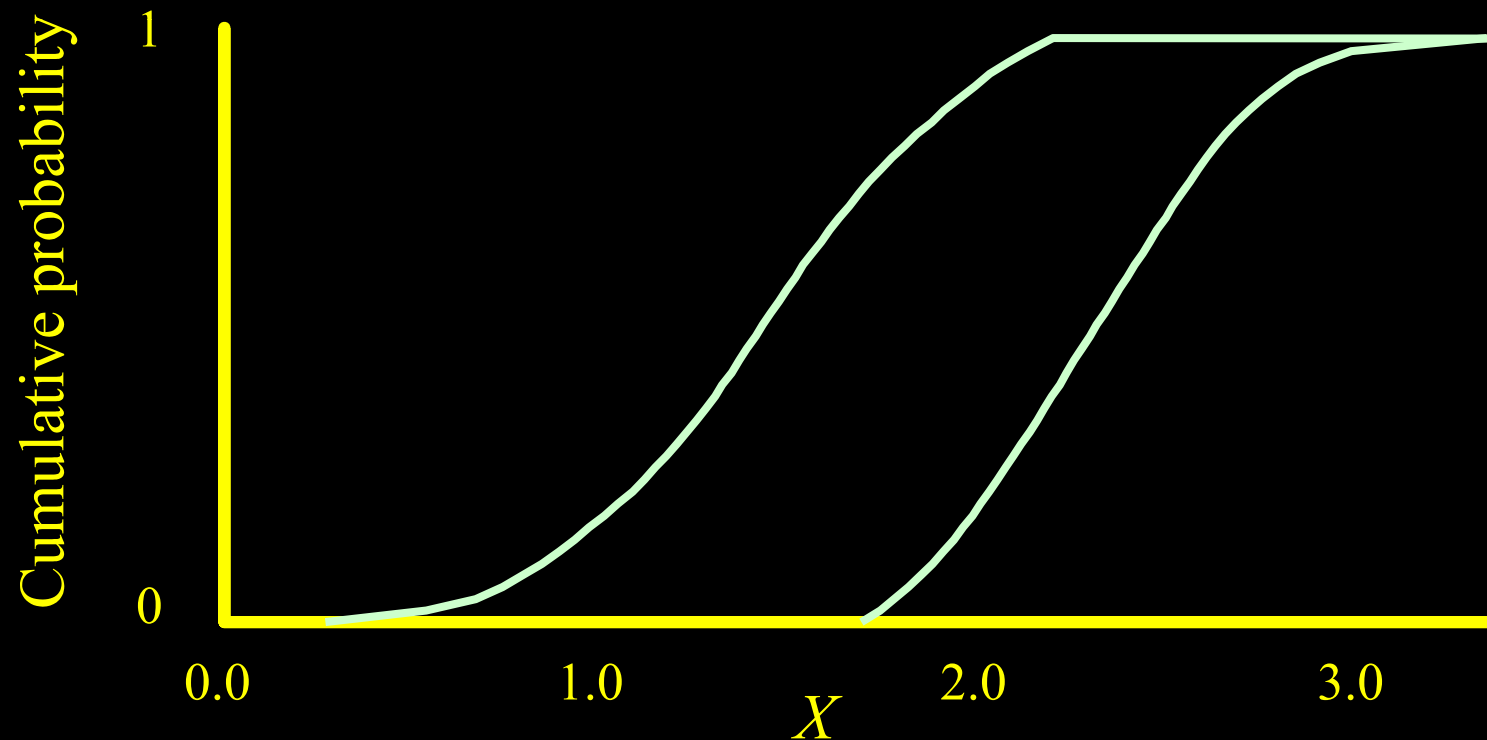
Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



Probability box (p-box)

Interval bounds on an cumulative distribution function (CDF)



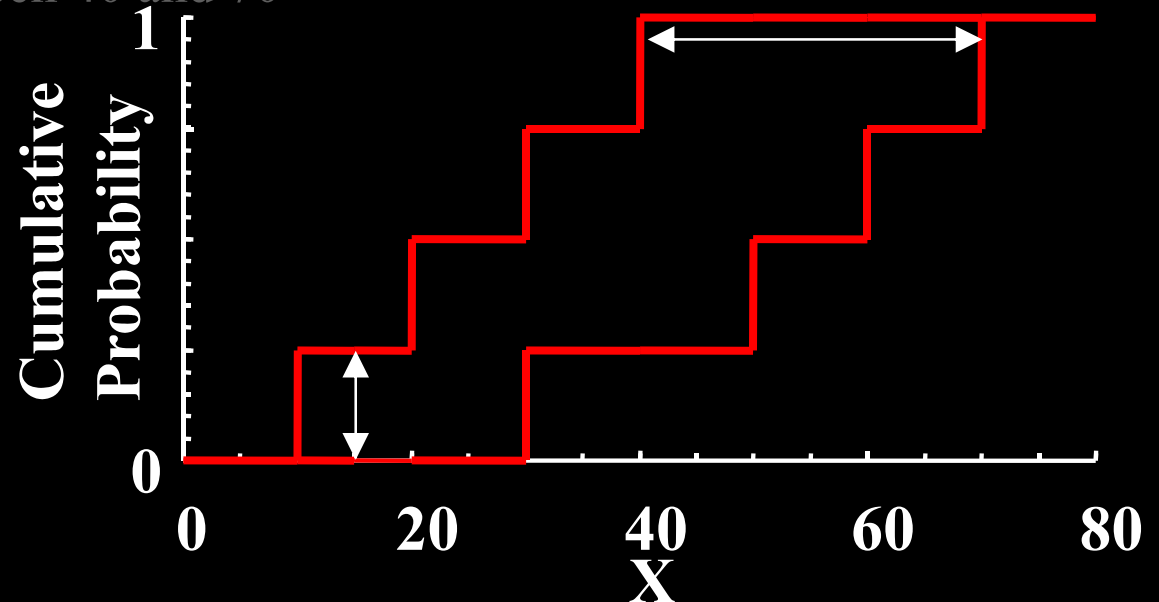
Duality

- Bounds on the probability at a value

Chance the value will be 15 or less is between 0 and 25%

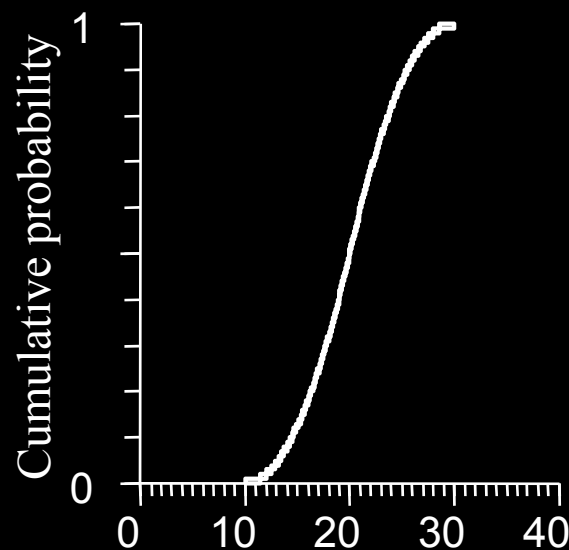
- Bounds on the value at a probability

95th percentile is between 40 and 70

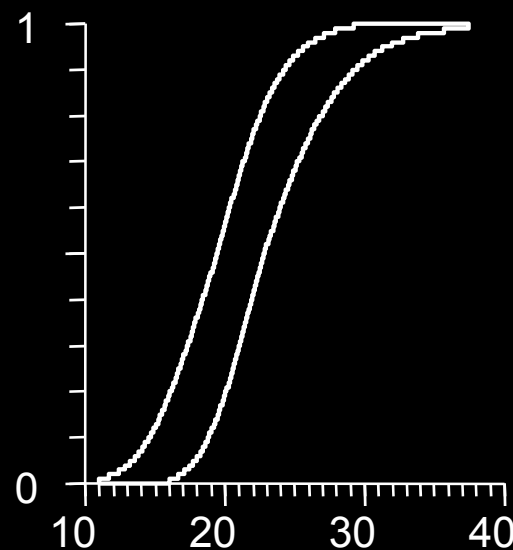


Generalizes an “uncertain number”

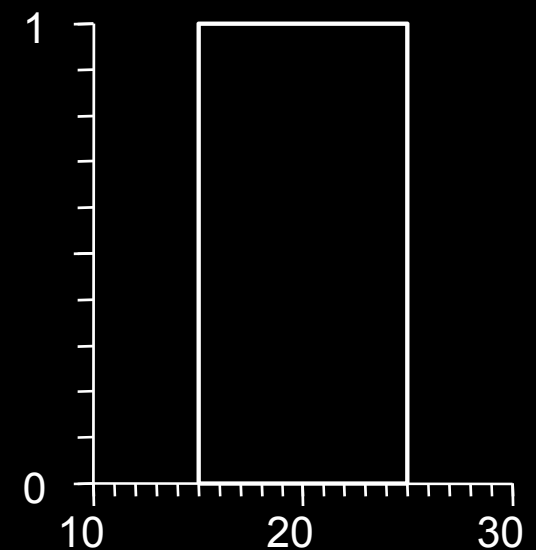
Probability
distribution



Probability
box



Interval



*Not a uniform
distribution*

Probability bounds analysis

- Marries intervals with probability theory
- Distinguishes variability and incertitude
- Solves many problems in uncertainty analysis
 - Input distributions unknown
 - Imperfectly known correlation and dependency
 - Large measurement error, censoring, small sample sizes
 - Model uncertainty

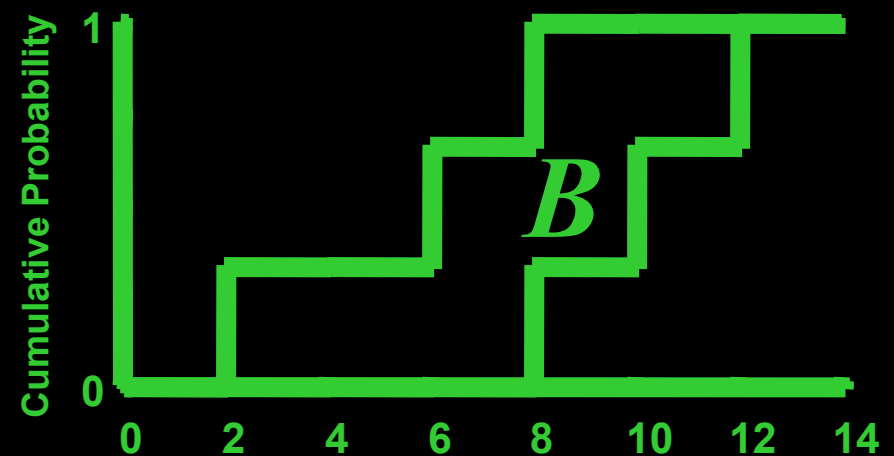
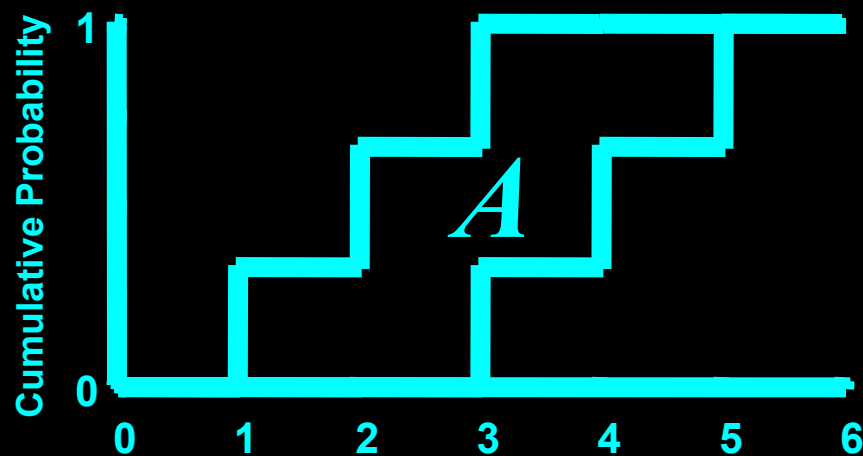
Calculations

- All standard mathematical operations
 - Arithmetic operations (+, −, ×, ÷, ^, min, max)
 - Logical operations (and, or, not, if, etc.)
 - Transformations (exp, ln, sin, tan, abs, sqrt, etc.)
 - Backcalculation (deconvolutions, updating)
 - Magnitude comparisons (<, ≤, >, ≥, ⊆)
 - Other operations (envelope, mixture, etc.)
- Faster than Monte Carlo
- Guaranteed to bounds answer
- Good solutions often easy to compute

Generalization of methods

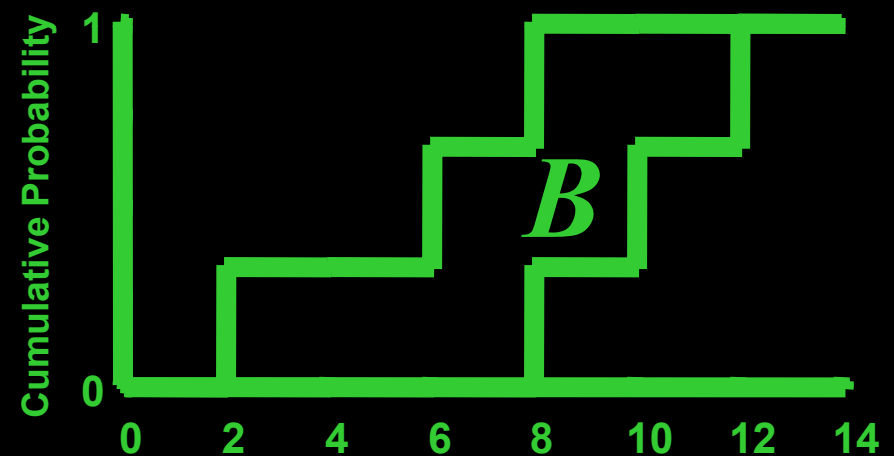
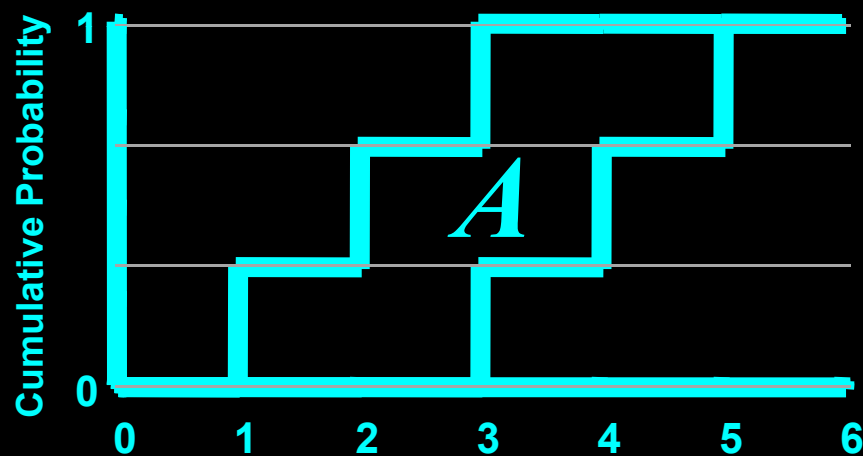
- When inputs are distributions, its answers conform with probability theory
- When inputs are intervals, it agrees with interval analysis

Probability bounds arithmetic



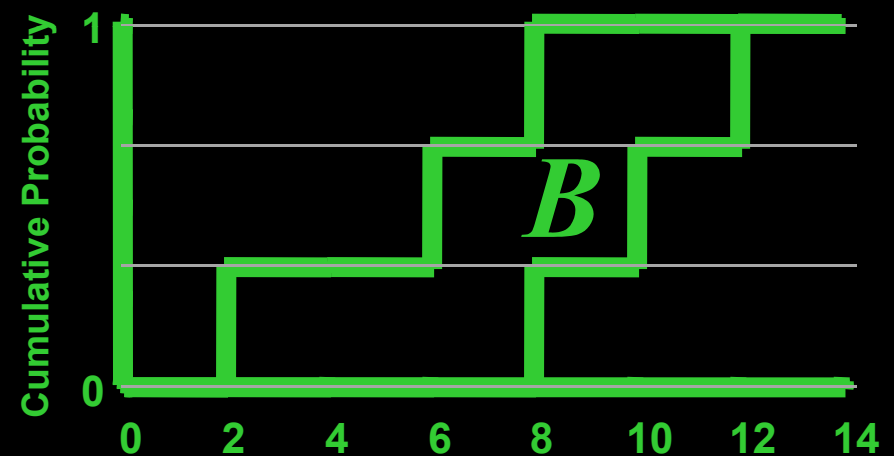
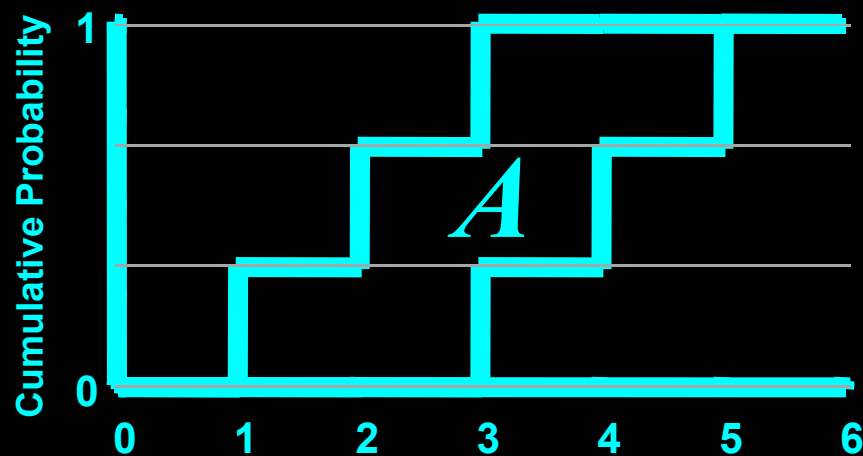
What's the sum of $A+B$?

Probability bounds arithmetic



What's the sum of $A+B$?

Probability bounds arithmetic



What's the sum of $A+B$?

Cartesian product

$A+B$

independence

$A \in [1,3]$
 $p_1 = 1/3$

$A \in [2,4]$
 $p_2 = 1/3$

$A \in [3,5]$
 $p_3 = 1/3$

$B \in [2,8]$
 $q_1 = 1/3$

$A+B \in [3,11]$
prob=1/9

$A+B \in [4,12]$
prob=1/9

$A+B \in [5,13]$
prob=1/9

$B \in [6,10]$
 $q_2 = 1/3$

$A+B \in [7,13]$
prob=1/9

$A+B \in [8,14]$
prob=1/9

$A+B \in [9,15]$
prob=1/9

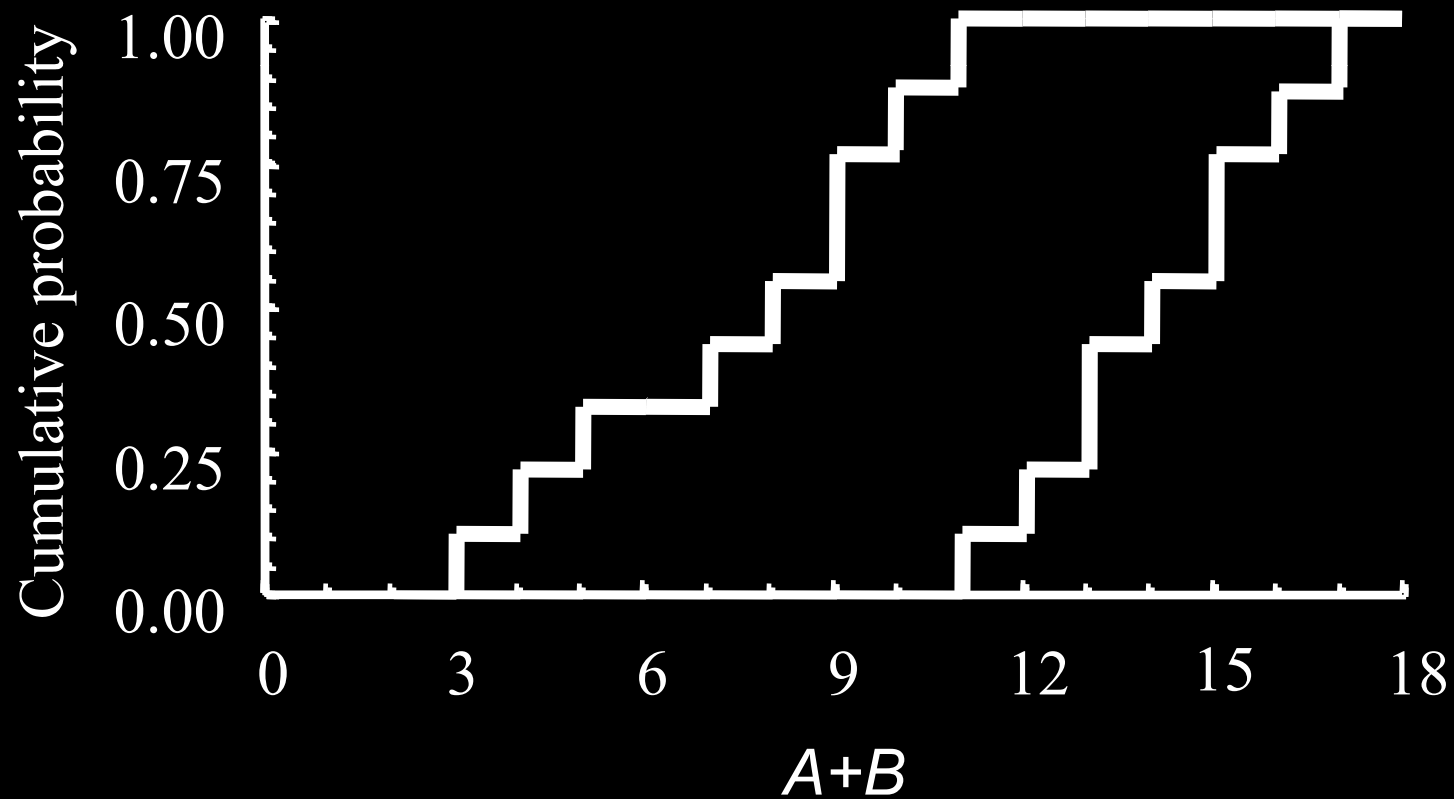
$B \in [8,12]$
 $q_3 = 1/3$

$A+B \in [9,15]$
prob=1/9

$A+B \in [10,16]$
prob=1/9

$A+B \in [11,17]$
prob=1/9

$A+B$ under independence



Where do we get p-boxes?

- Assumption
- Modeling
- Robust Bayesian analysis
- Constraint propagation
- Data with incertitude
 - Measurement error
 - Sampling error
 - Censoring

Data of poor repute

- Sometimes measurements are *intervals*
- Statisticians often ignore interval uncertainty, or model it as a probability distribution
- “Interval uncertainty doesn’t exist in real life”

Incertitude is common in data

- Periodic observations

When did the fish in my aquarium die during the night?

- Plus-or-minus measurement uncertainties

Coarse measurements, measurements from digital readouts

- Non-detects and data censoring

Chemical detection limits, studies prematurely terminated

- Privacy requirements

Epidemiological or medical information, census data

- Theoretical constraints

Concentrations, solubilities, probabilities, survival rates

- Bounding studies

Presumed or hypothetical limits in what-if calculations

A tale of two data sets

Skinny data

[1.00, 1.52]

[2.68, 2.98]

[7.52, 7.67]

[7.73, 8.35]

[9.44, 9.99]

[3.66, 4.58]

Puffy data

[3.5, 6.4]

[6.9, 8.8]

[6.1, 8.4]

[2.8, 6.7]

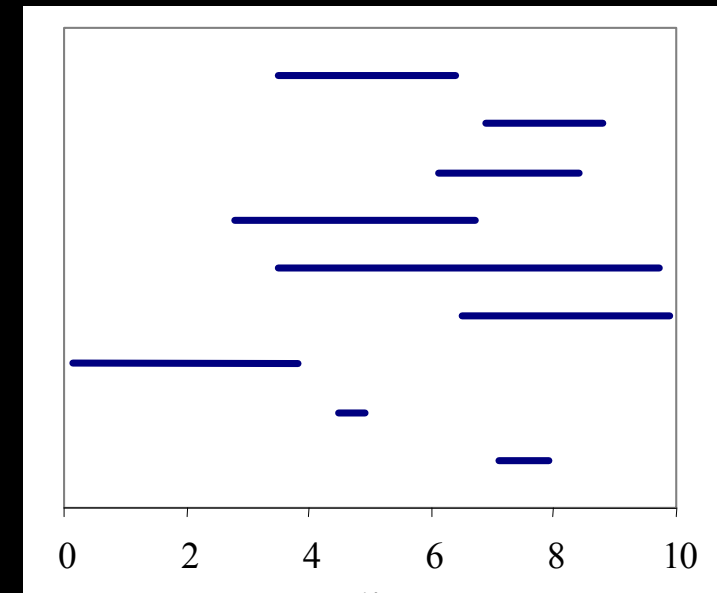
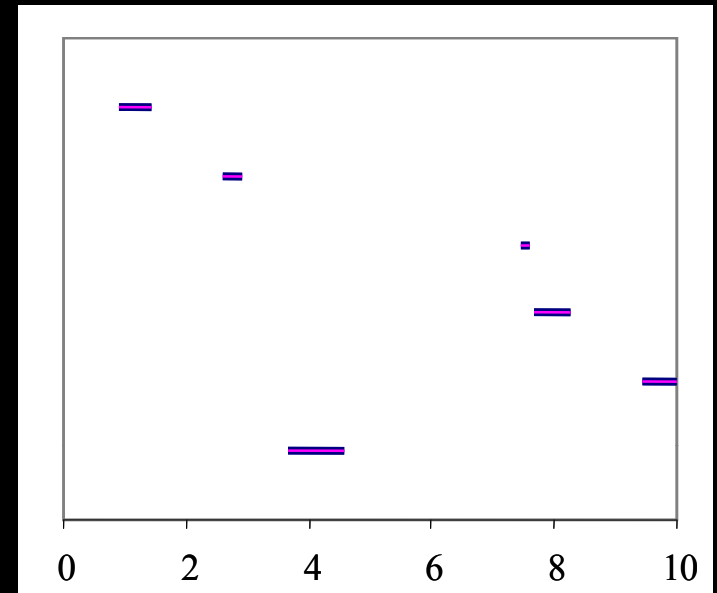
[3.5, 9.7]

[6.5, 9.9]

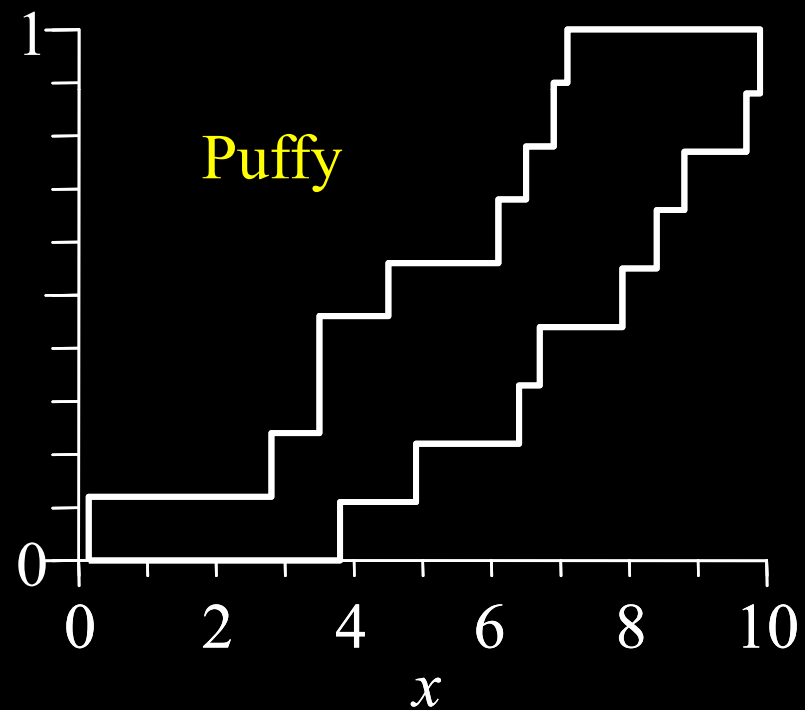
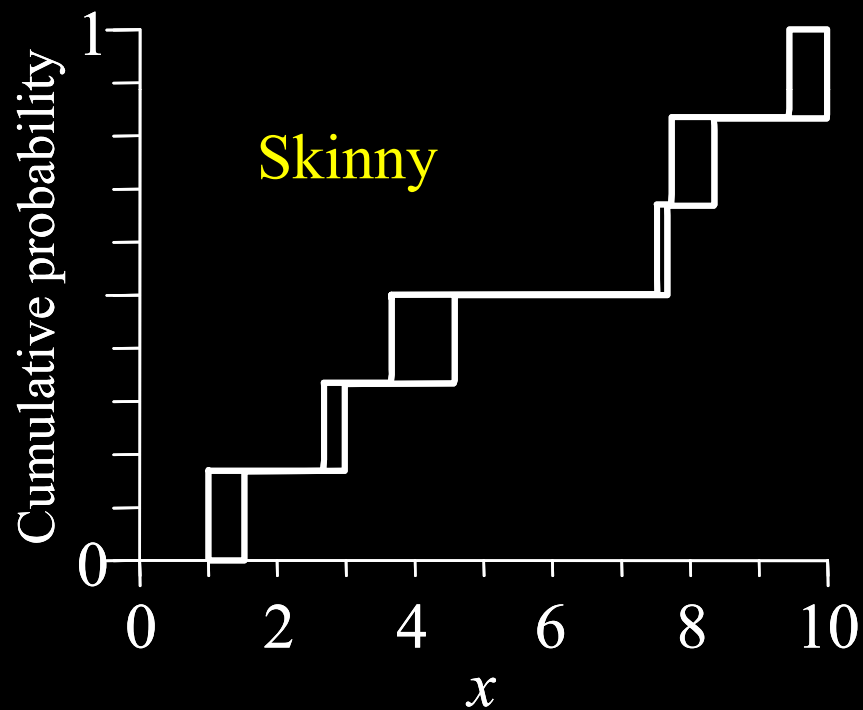
[0.15, 3.8]

[4.5, 4.9]

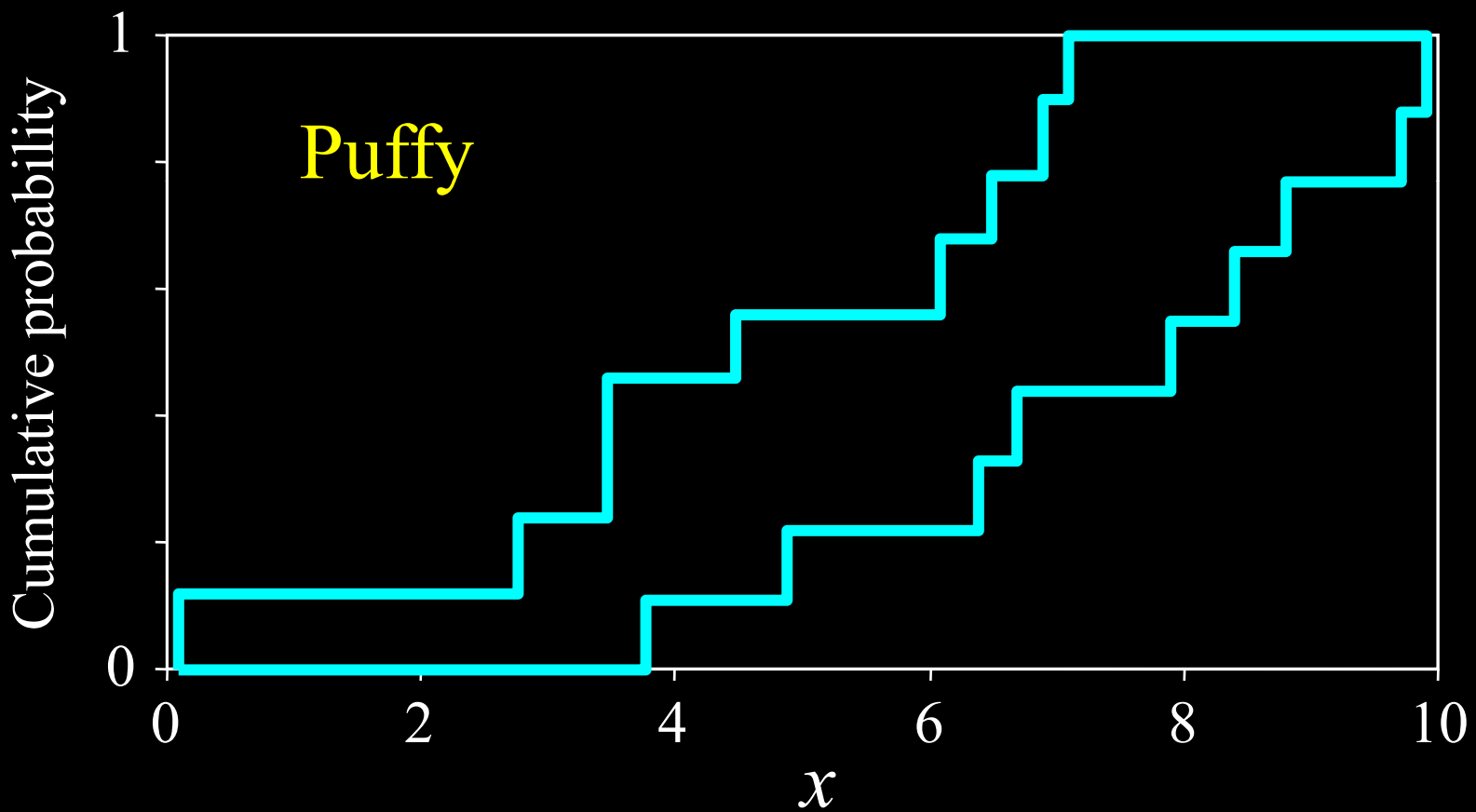
[7.1, 7.9]



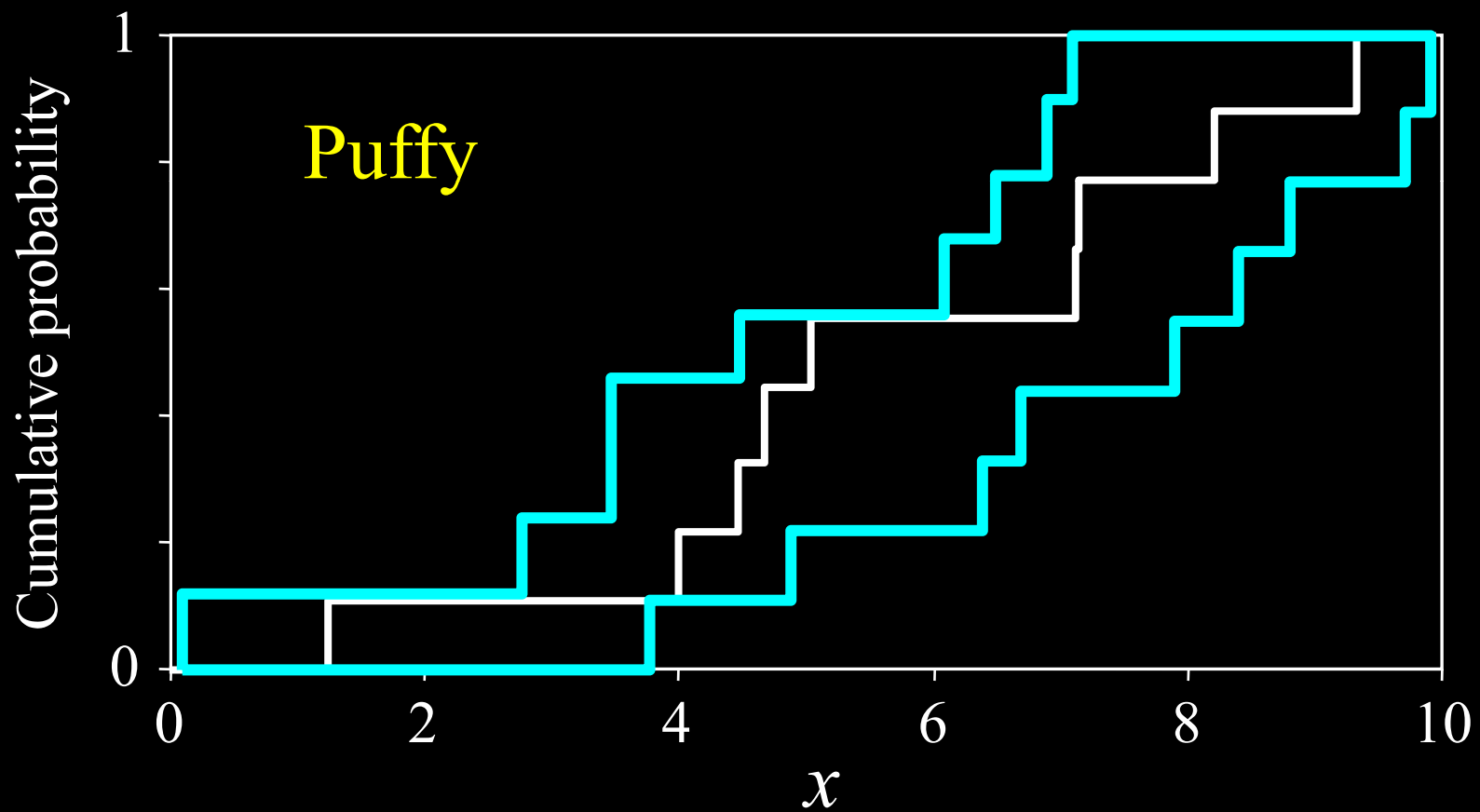
Empirical distributions



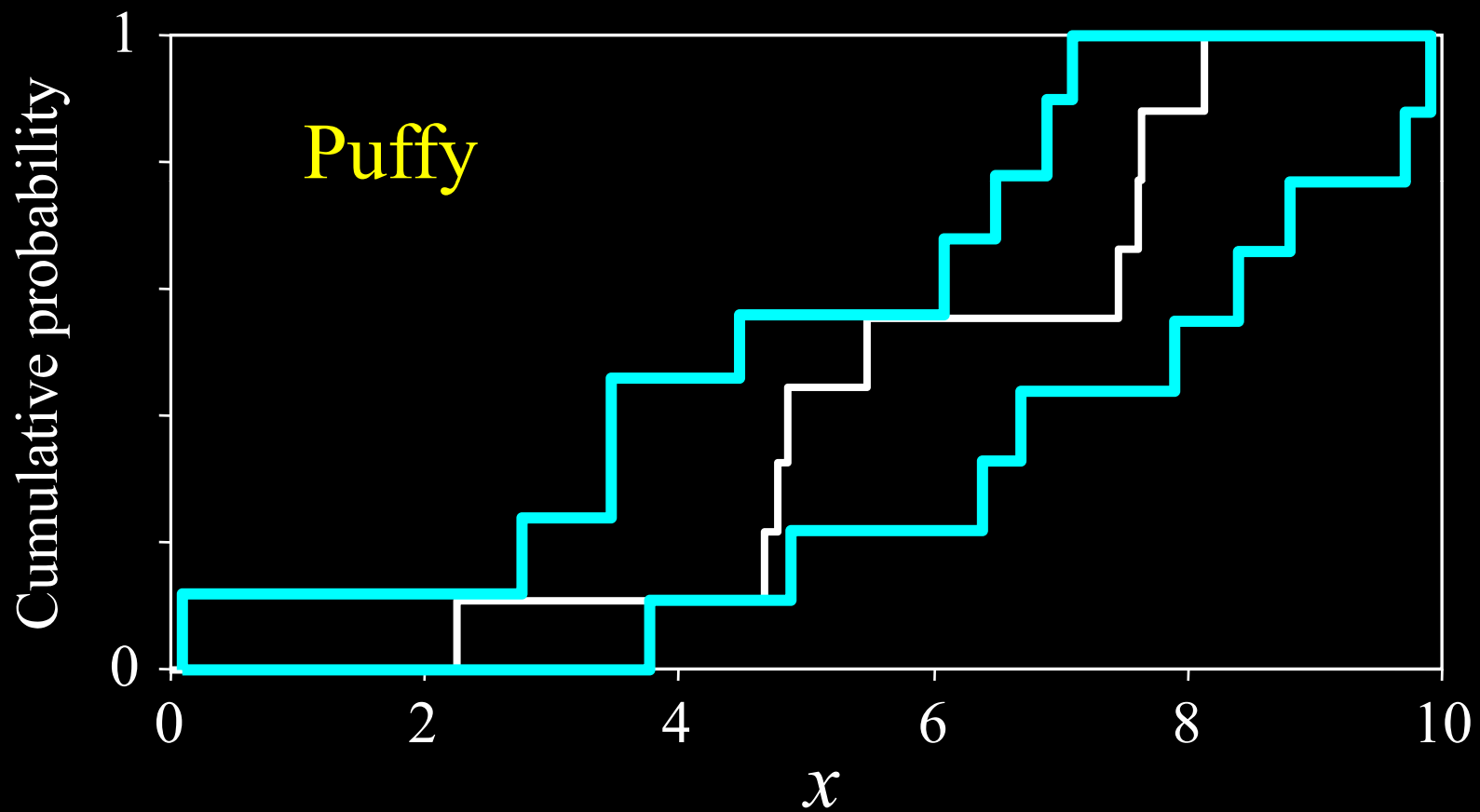
Many possible precise data sets



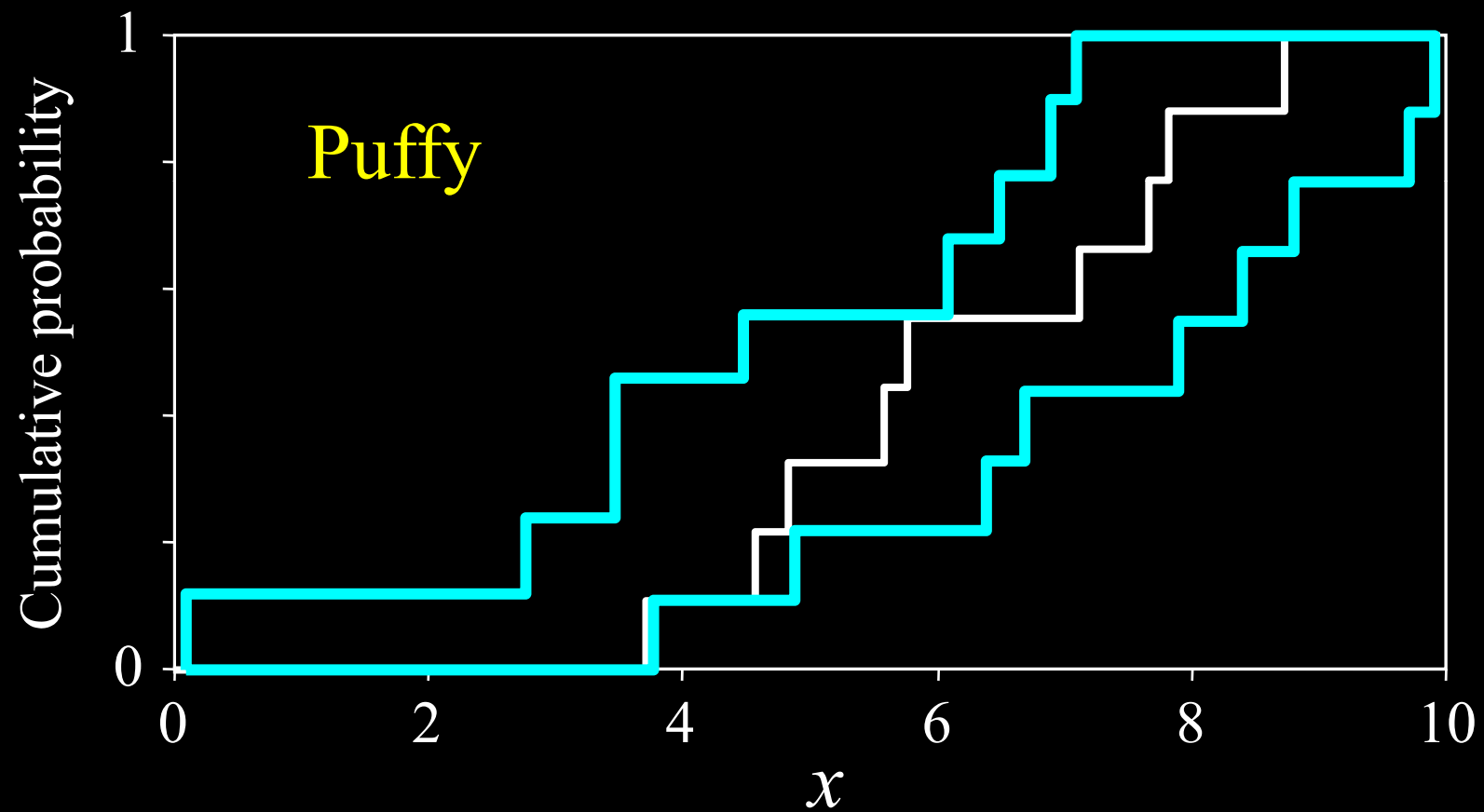
Many possible precise data sets



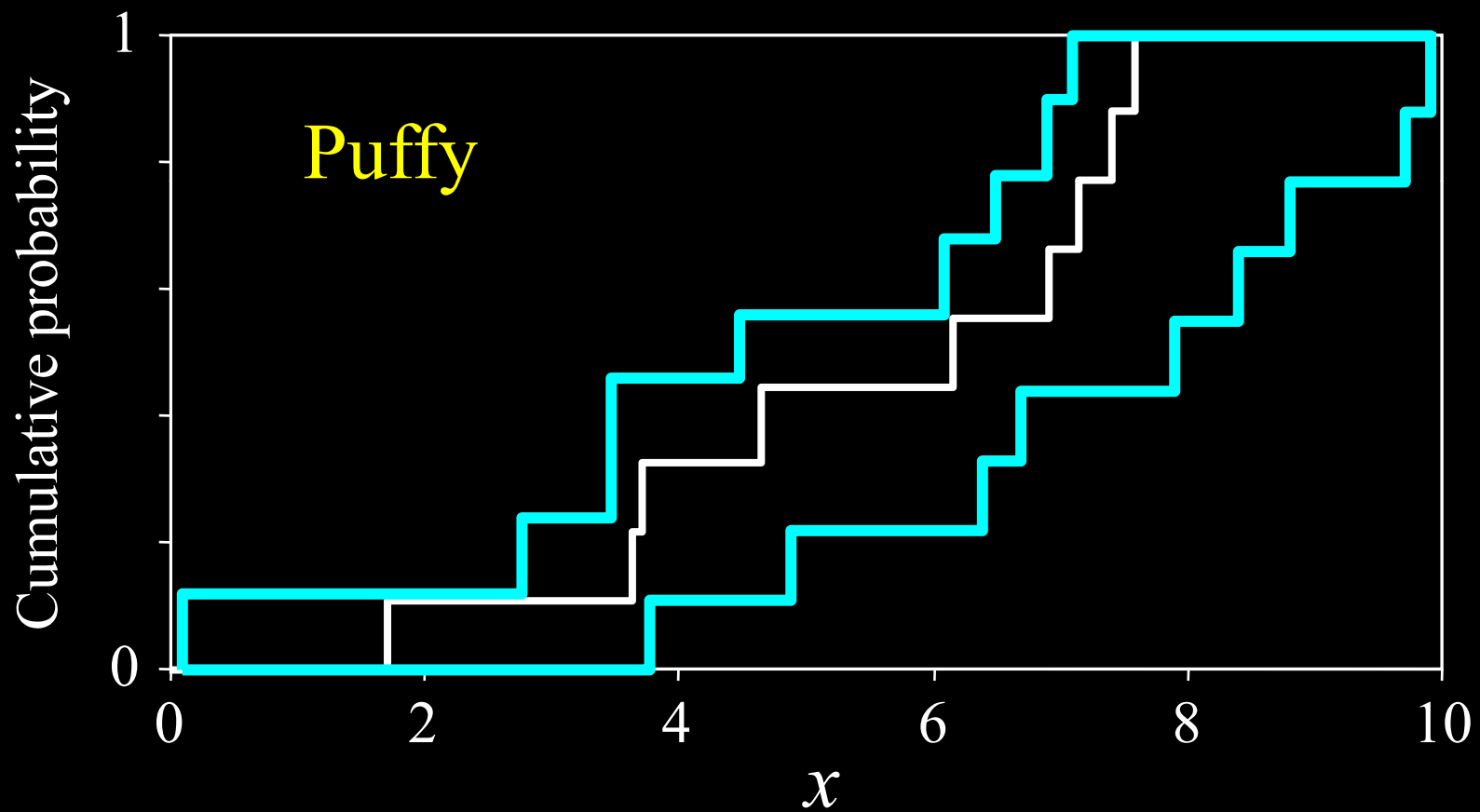
Many possible precise data sets



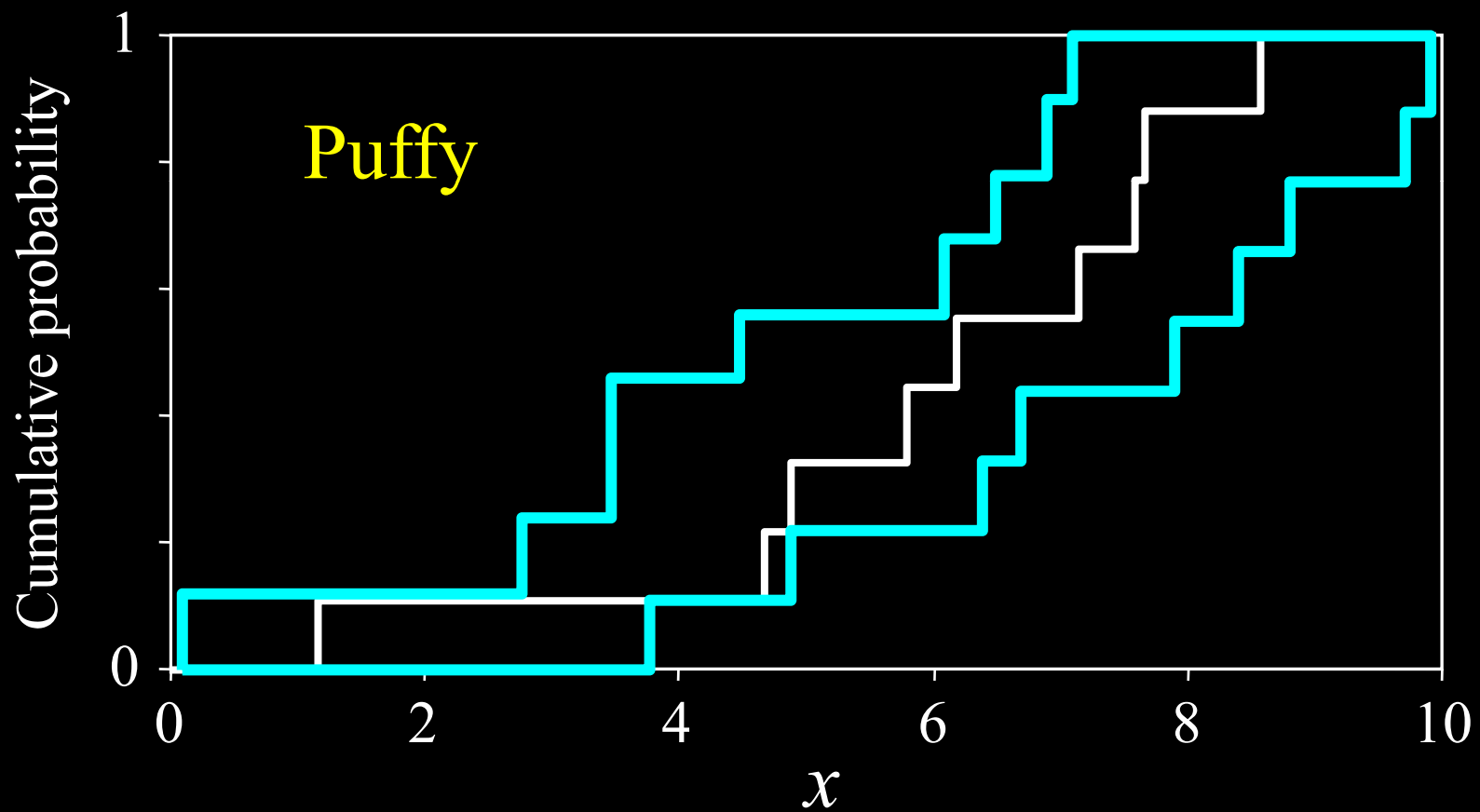
Many possible precise data sets



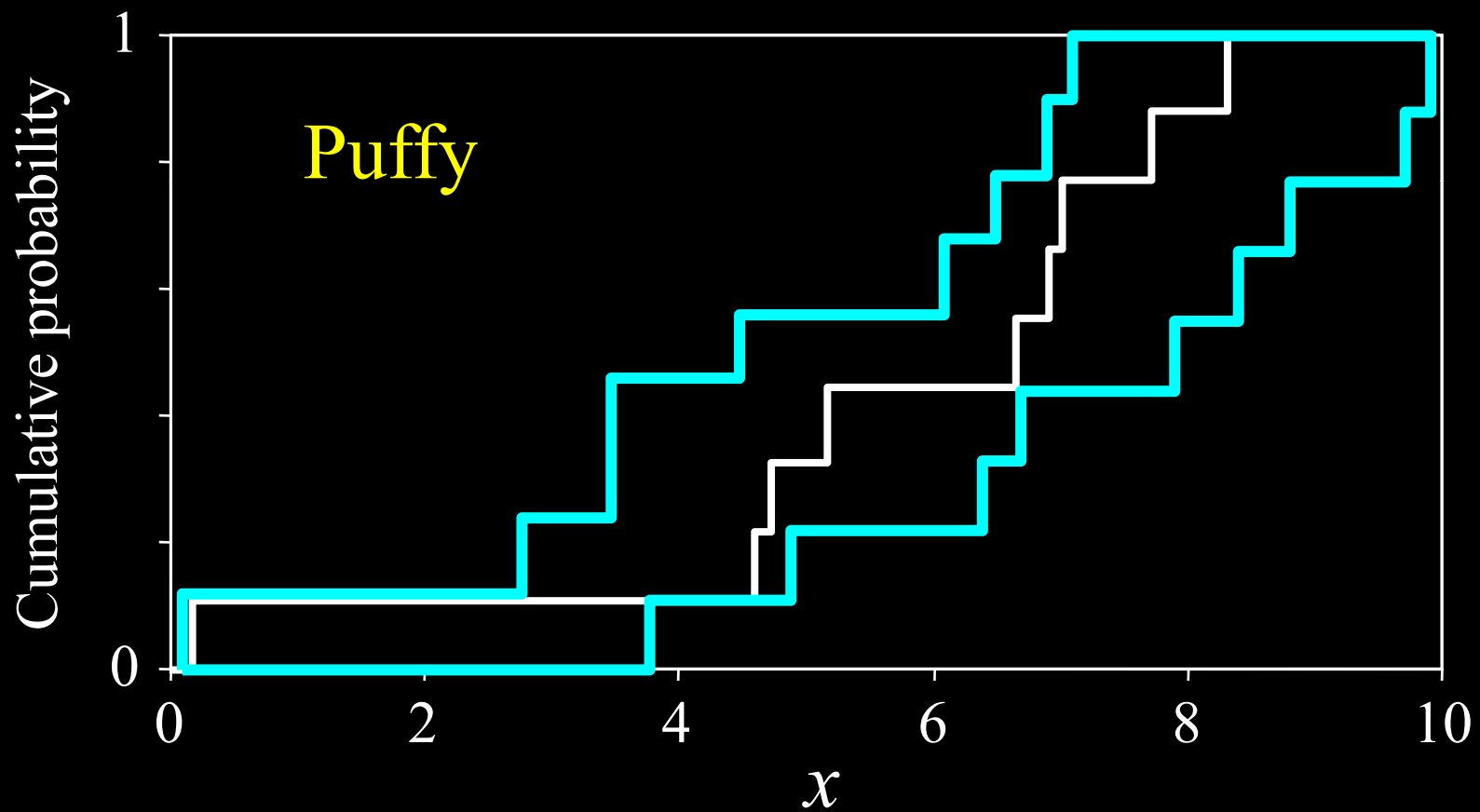
Many possible precise data sets



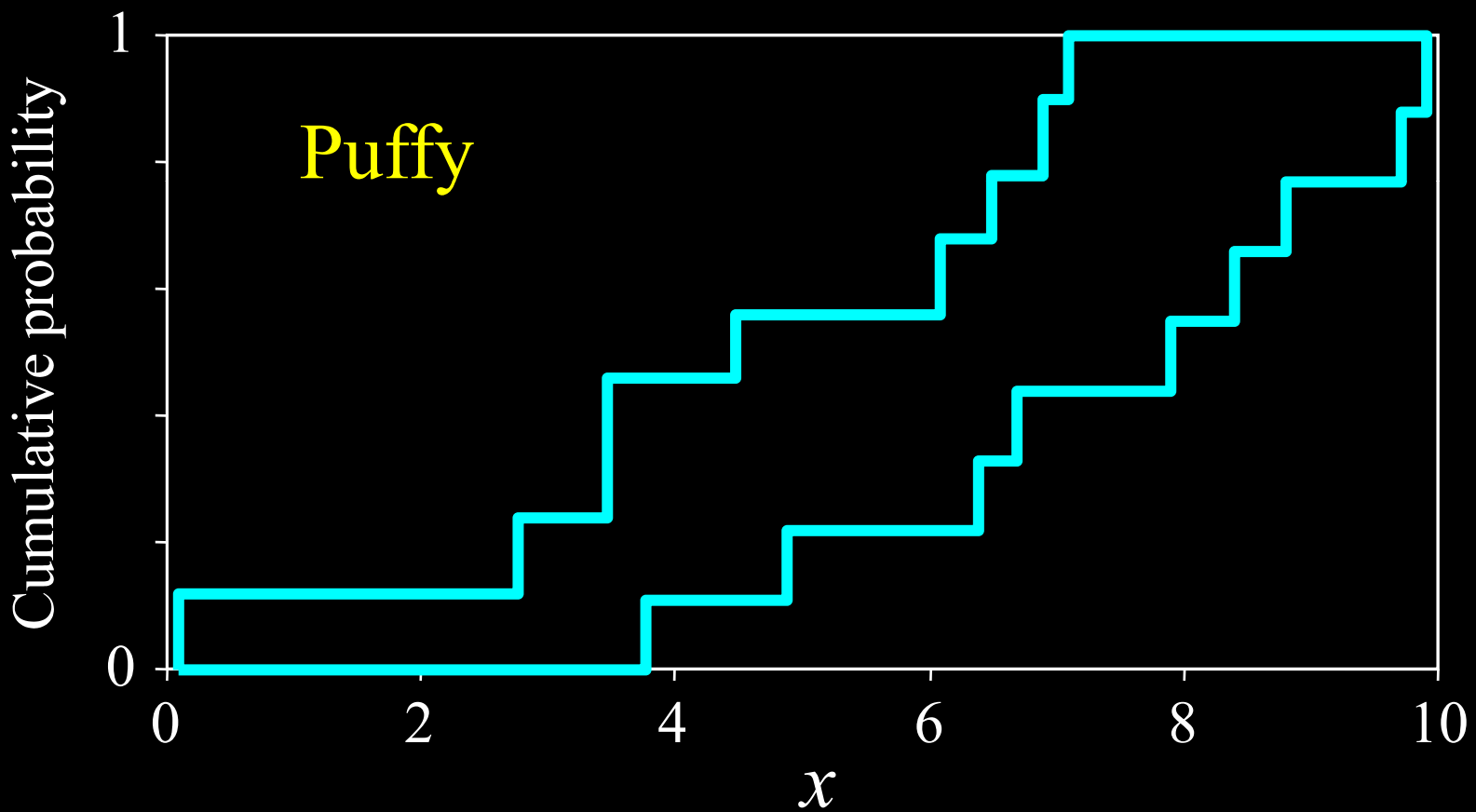
Many possible precise data sets



Many possible precise data sets



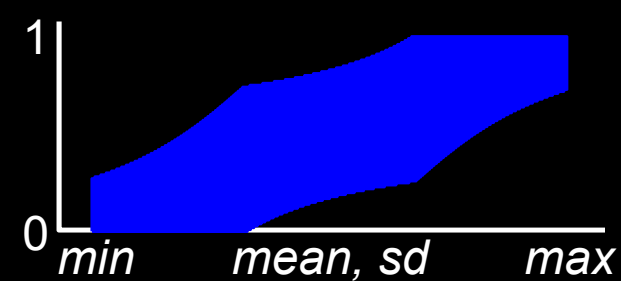
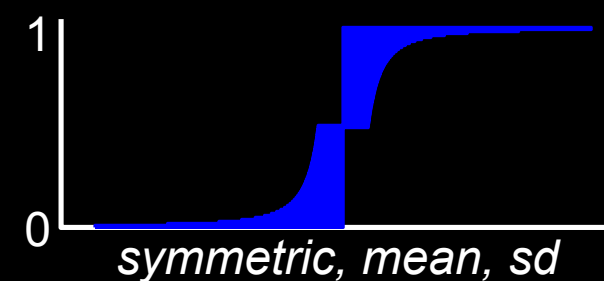
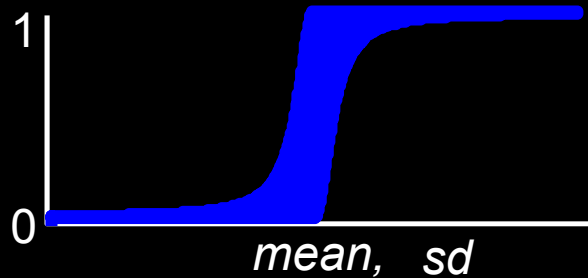
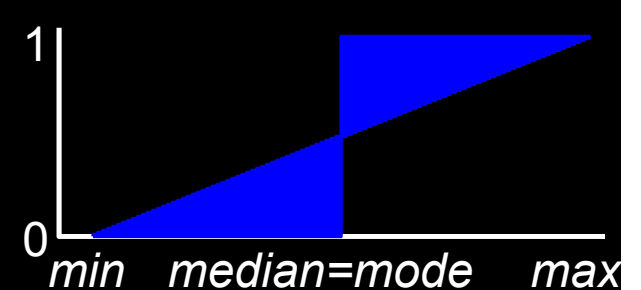
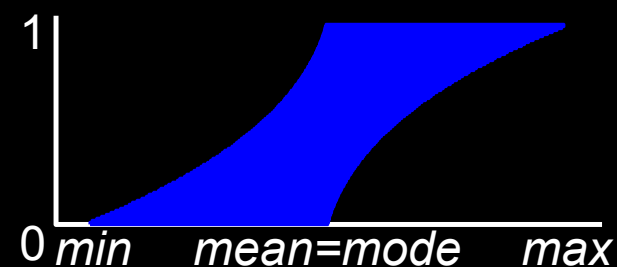
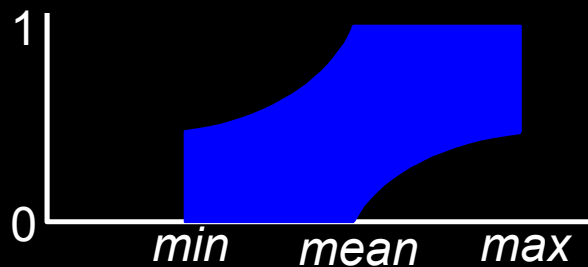
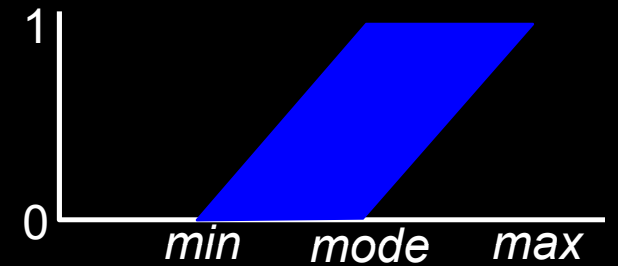
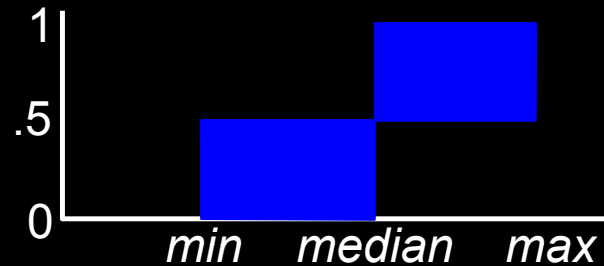
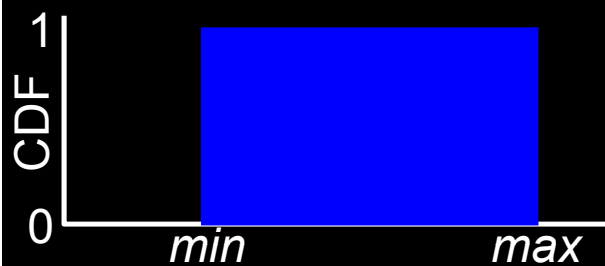
Many possible precise data sets



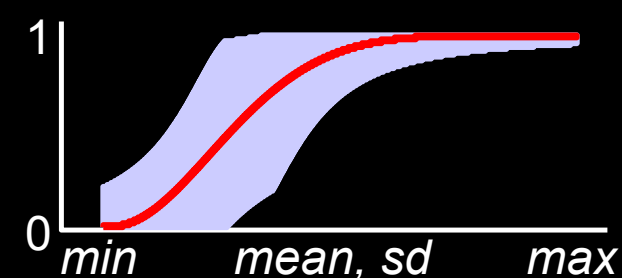
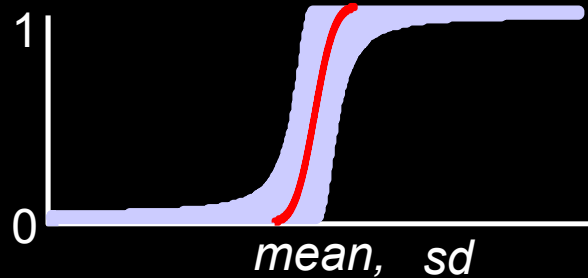
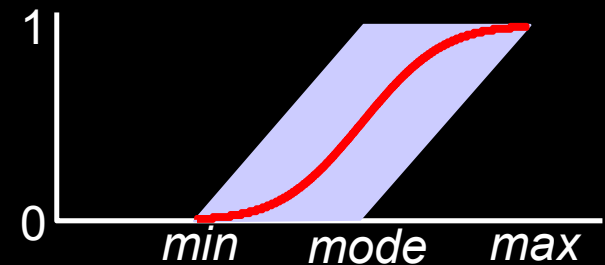
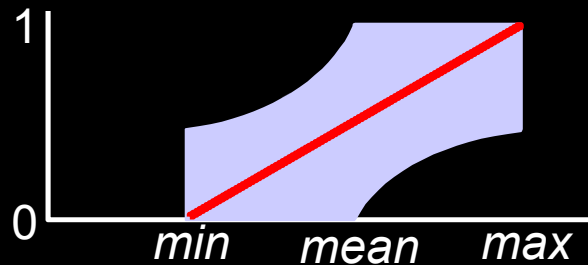
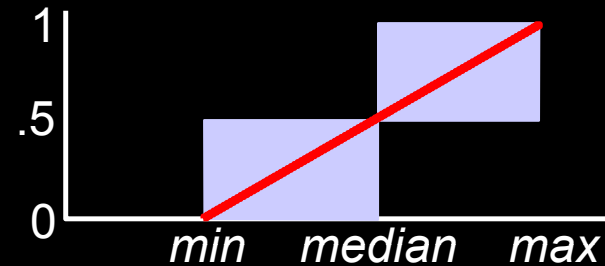
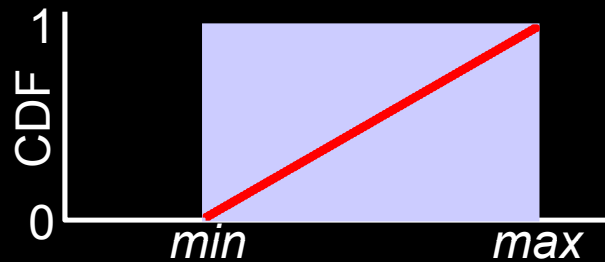
Interval statistics

- Statistics for data sets that contain intervals
- Some are easy to compute
 - Empirical distribution, mean, median, percentiles, etc.
- Some are tricky, but easy for a computer
 - Variance, upper confidence limit, correlation, etc.
- Tradeoff between more versus better data
- Review just published as a Sandia report

Constraint propagation



Maximum entropy erases uncertainty



Example: PCBs and duck hunters

Location: Massachusetts and Connecticut

Receptor: Adult human hunters of waterfowl

Contaminant: PCBs (polychlorinated biphenyls)

Exposure route: dietary consumption of
contaminated waterfowl

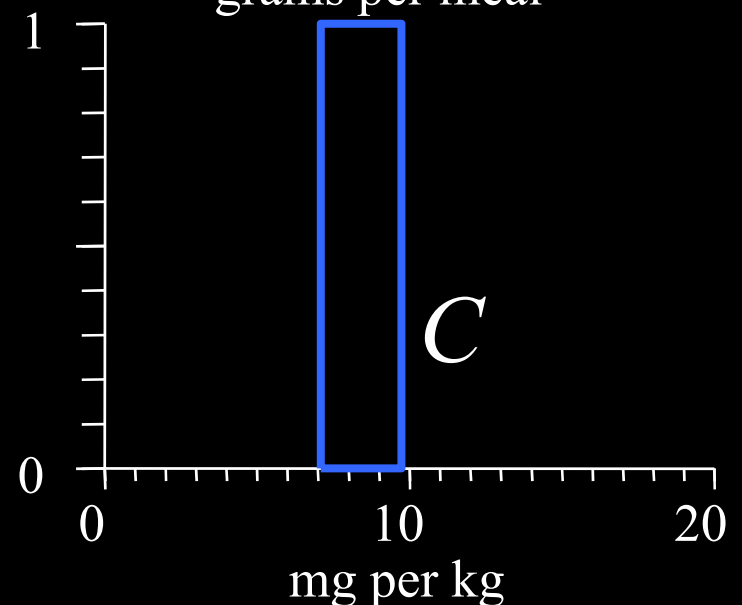
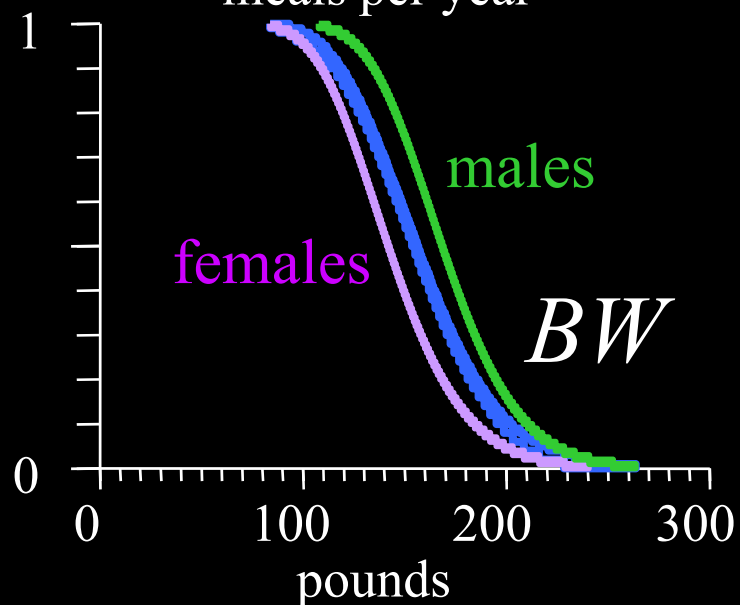
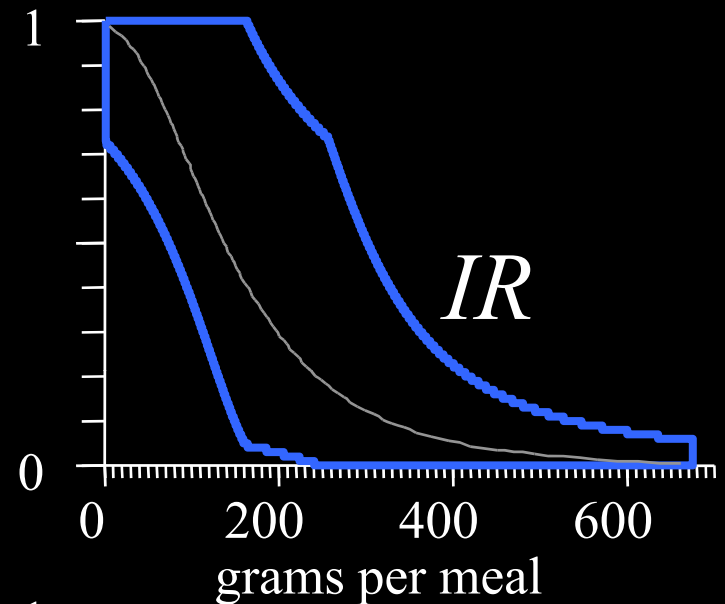
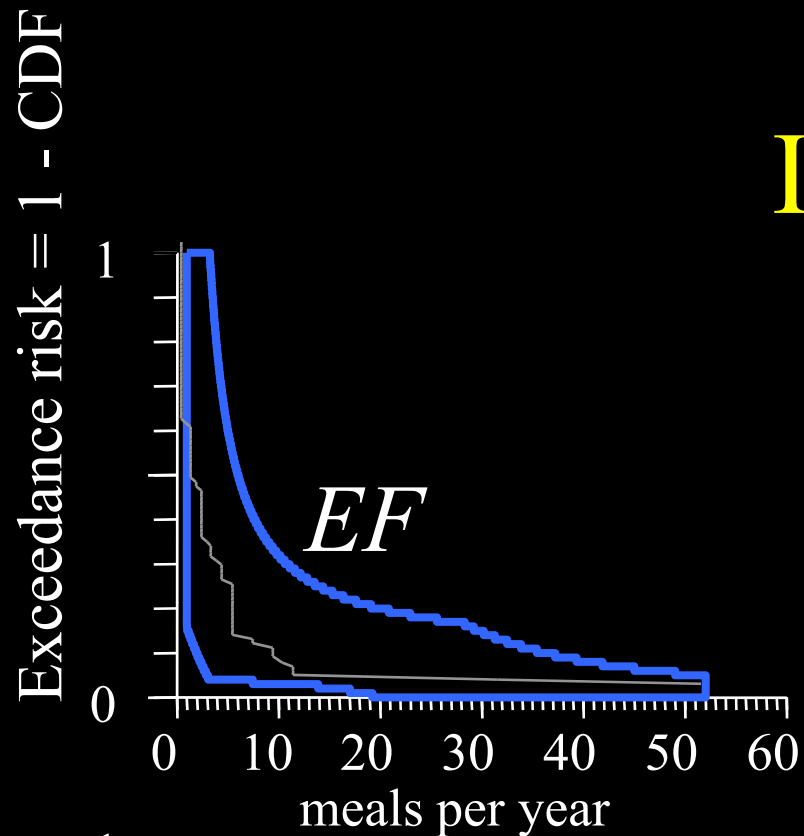
Based on the assessment for non-cancer risks from PCB to adult hunters who consume contaminated waterfowl described in *Human Health Risk Assessment: GE/Housatonic River Site: Rest of River*, Volume IV, DCN: GE-031203-ABMP, April 2003, Weston Solutions (West Chester, Pennsylvania), Avatar Environmental (Exton, Pennsylvania), and Applied Biomathematics (Setauket, New York).

Hazard quotient

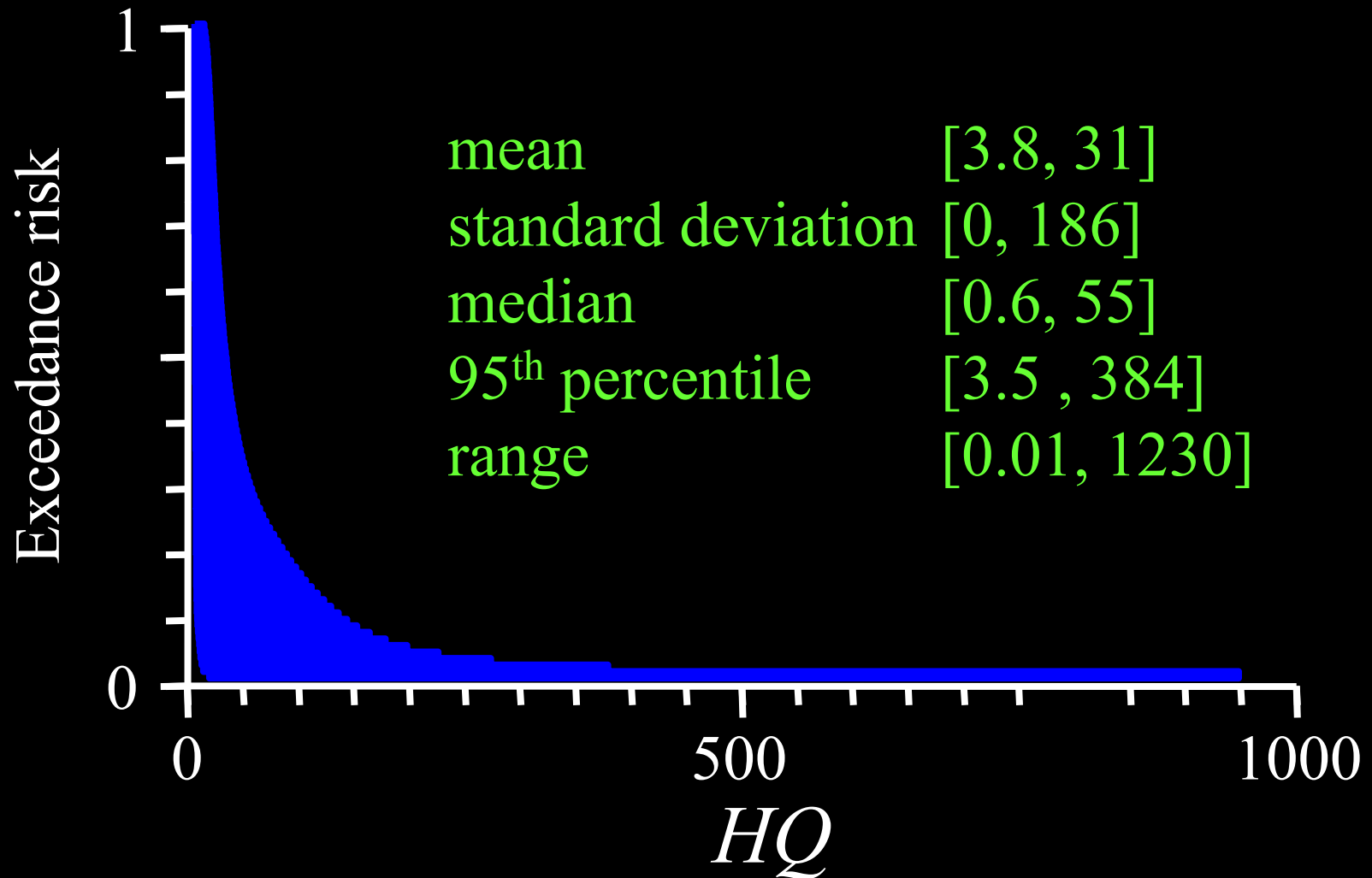
$$HQ = \frac{EF \times IR \times C \times (1 - LOSS)}{AT \times BW \times RfD}$$

$EF = \text{mmms}(1, 52, 5.4, 10)$ meals per year // exposure frequency, censored data, $n = 23$
 $IR = \text{mmms}(1.5, 675, 188, 113)$ grams per meal // poultry ingestion rate from EPA's *EFH*
 $C = [7.1, 9.73]$ mg per kg // exposure point (mean) concentration
 $LOSS = 0$ // loss due to cooking
 $AT = 365.25$ days per year // averaging time (not just units conversion)
 $BW = \text{mixture}(BW_{\text{female}}, BW_{\text{male}})$ // Brainard and Burmaster (1992)
 $BW_{\text{male}} = \text{lognormal}(171, 30)$ pounds // adult male $n = 9,983$
 $BW_{\text{female}} = \text{lognormal}(145, 30)$ pounds // adult female $n = 10,339$
 $RfD = 0.00002$ mg per kg per day // reference dose considered tolerable

Inputs



Results



Rigorousness

- “Automatically verified calculations”
- The computations are guaranteed to enclose the true results (so long as the inputs do)
- You can still be wrong, but the *method* won't be the reason if you are

How to use the results

When uncertainty makes no difference

(because results are so clear), bounding gives confidence in the reliability of the decision

When uncertainty obscures the decision

- (i) use results to identify inputs to study better, or
- (ii) use other criteria within probability bounds

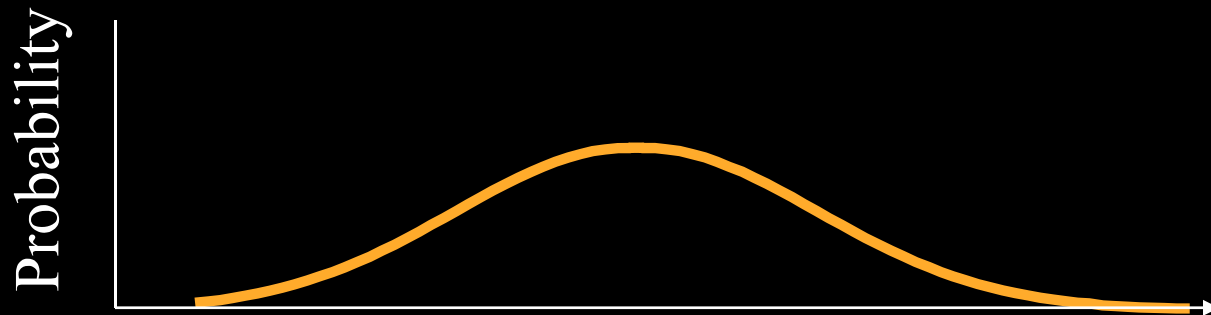
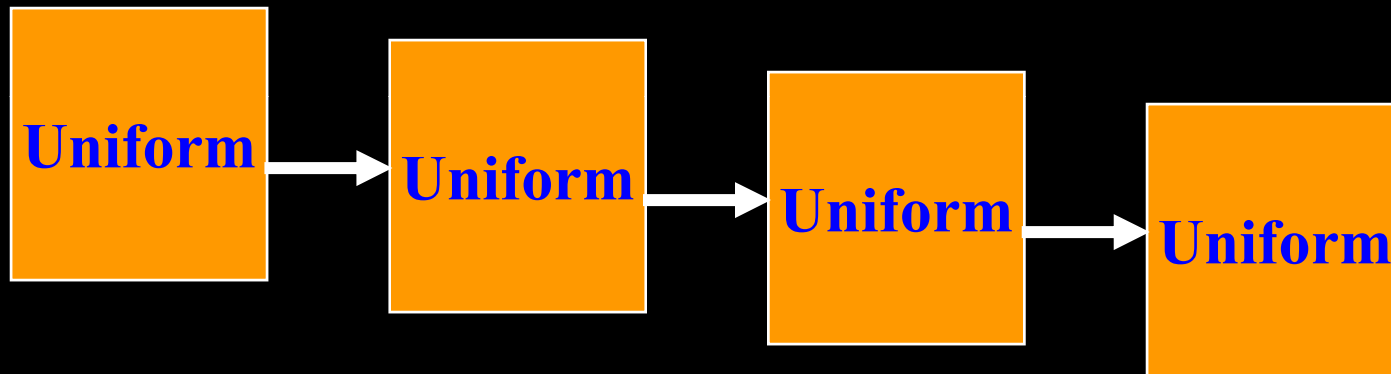
Can uncertainty swamp the answer?

- Sure, if uncertainty is huge
- This *should* happen (it's not “unhelpful”)
- If you think the bounds are too wide, then put in whatever information is missing
- If there isn't any such information, do you want to *mislead* your readers?

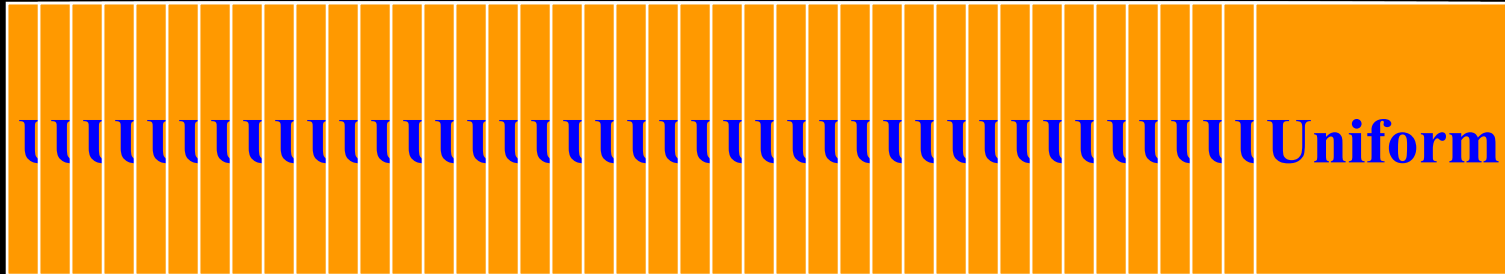
Monte Carlo is problematic

- Probability doesn't accumulate gross uncertainty in an intuitive way
- Precision of the answer (measured as cv) depends strongly on the number of inputs
- The more inputs, the tighter the answer, irrespective of the distribution shape

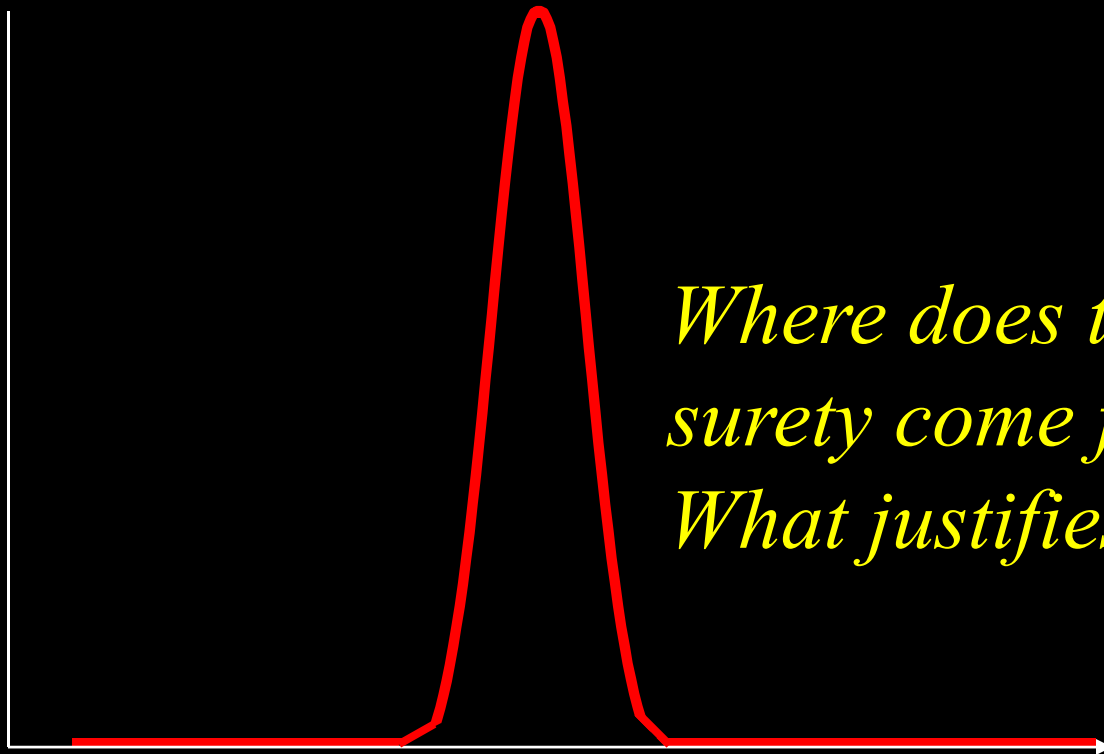
A few grossly uncertain inputs



A lot of grossly uncertain inputs...



Probability



*Where does this
surety come from?
What justifies it?*

Smoke and mirrors certainty

- P-boxes give a vacuous answer if all you provide are vacuous inputs
- Conventional probability theory, at least as it's naively applied, seems to manufacture certainty out of nothing
- This is why some critics say probabilistic risk analyses are “smoke and mirrors”

Uncertainty about distribution shape

- Maximum entropy hides uncertainty
- Can't propagate with sensitivity analysis
- Bounding best, but must reflect all available information

Dependence

Dependence

- Not all variables are independent
 - Body size and skin surface area
 - “Common-cause” variables
- Known dependencies should be modeled
- What can we do when we don't know them?

How to do other dependencies?

- Independent
- Perfect (comonotonic)
- Opposite (countermonotonic)

Perfect dependence

$A+B$ perfect positive	$A \in [1,3]$ $p_1 = 1/3$	$A \in [2,4]$ $p_2 = 1/3$	$A \in [3,5]$ $p_3 = 1/3$
$B \in [2,8]$ $q_1 = 1/3$	$A+B \in [3,11]$ prob=1/3	$A+B \in [4,12]$ prob=0	$A+B \in [5,13]$ prob=0
$B \in [6,10]$ $q_2 = 1/3$	$A+B \in [7,13]$ prob=0	$A+B \in [8,14]$ prob=1/3	$A+B \in [9,15]$ prob=0
$B \in [8,12]$ $q_3 = 1/3$	$A+B \in [9,15]$ prob=0	$A+B \in [10,16]$ prob=0	$A+B \in [11,17]$ prob=1/3

Perfect dependence

$A+B$ perfect positive	$A \in [1,3]$ $p_1 = 1/3$	$A \in [2,4]$ $p_2 = 1/3$	$A \in [3,5]$ $p_3 = 1/3$
$B \in [2,8]$ $q_1 = 1/3$	$A+B \in [3,11]$ prob=1/3	$A+B \in [4,12]$ prob=0	$A+B \in [5,13]$ prob=0
$B \in [6,10]$ $q_2 = 1/3$	$A+B \in [7,13]$ prob=0	$A+B \in [8,14]$ prob=1/3	$A+B \in [9,15]$ prob=0
$B \in [8,12]$ $q_3 = 1/3$	$A+B \in [9,15]$ prob=0	$A+B \in [10,16]$ prob=0	$A+B \in [11,17]$ prob=1/3

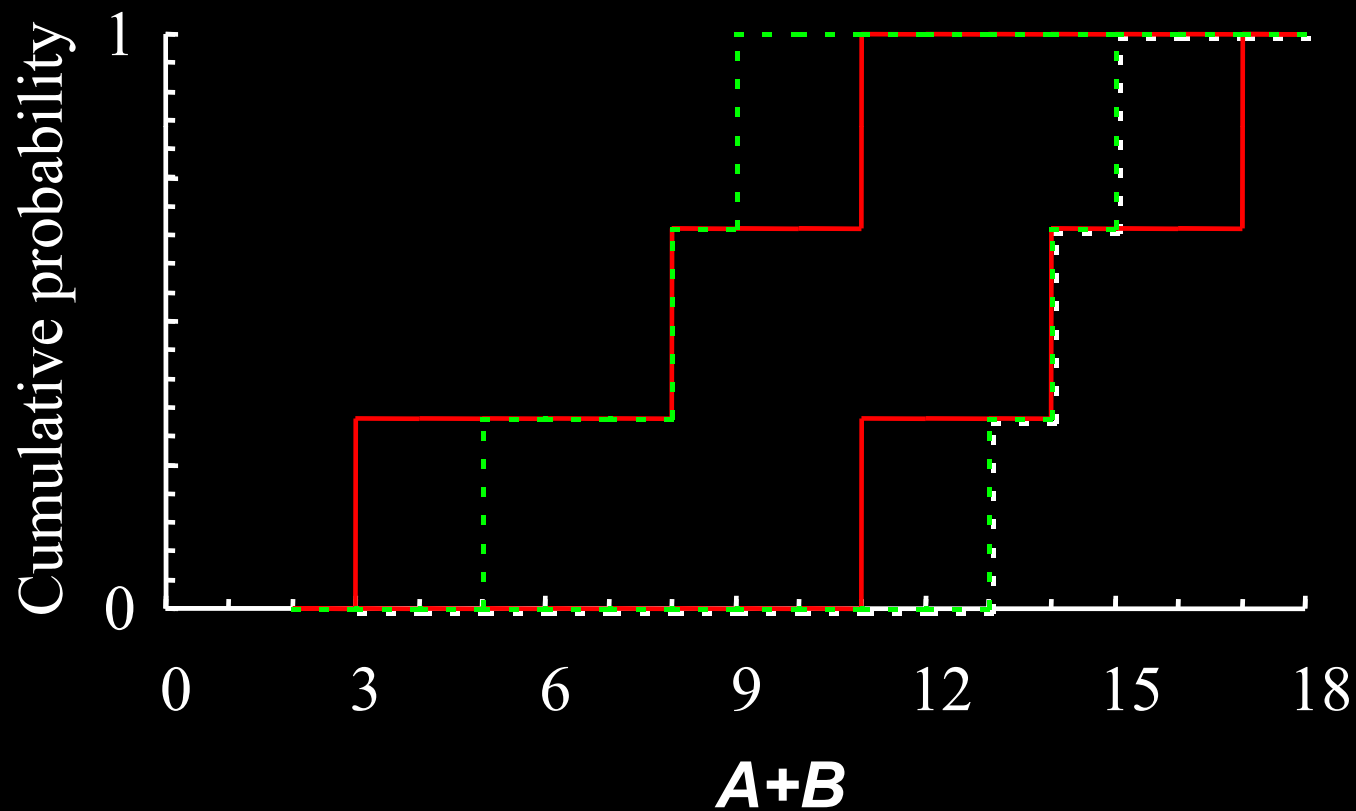
Opposite dependence

$A+B$ opposite positive	$A \in [1,3]$ $p_1 = 1/3$	$A \in [2,4]$ $p_2 = 1/3$	$A \in [3,5]$ $p_3 = 1/3$
$B \in [2,8]$ $q_1 = 1/3$	$A+B \in [3,11]$ prob=0	$A+B \in [4,12]$ prob=0	$A+B \in [5,13]$ prob=1/3
$B \in [6,10]$ $q_2 = 1/3$	$A+B \in [7,13]$ prob=0	$A+B \in [8,14]$ prob=1/3	$A+B \in [9,15]$ prob=0
$B \in [8,12]$ $q_3 = 1/3$	$A+B \in [9,15]$ prob= 1/3	$A+B \in [10,16]$ prob=0	$A+B \in [11,17]$ prob=0

Opposite dependence

$A+B$ opposite positive	$A \in [1,3]$ $p_1 = 1/3$	$A \in [2,4]$ $p_2 = 1/3$	$A \in [3,5]$ $p_3 = 1/3$
$B \in [2,8]$ $q_1 = 1/3$	$A+B \in [3,11]$ prob=0	$A+B \in [4,12]$ prob=0	$A+B \in [5,13]$ prob=1/3
$B \in [6,10]$ $q_2 = 1/3$	$A+B \in [7,13]$ prob=0	$A+B \in [8,14]$ prob=1/3	$A+B \in [9,15]$ prob=0
$B \in [8,12]$ $q_3 = 1/3$	$A+B \in [9,15]$ prob= 1/3	$A+B \in [10,16]$ prob=0	$A+B \in [11,17]$ prob=0

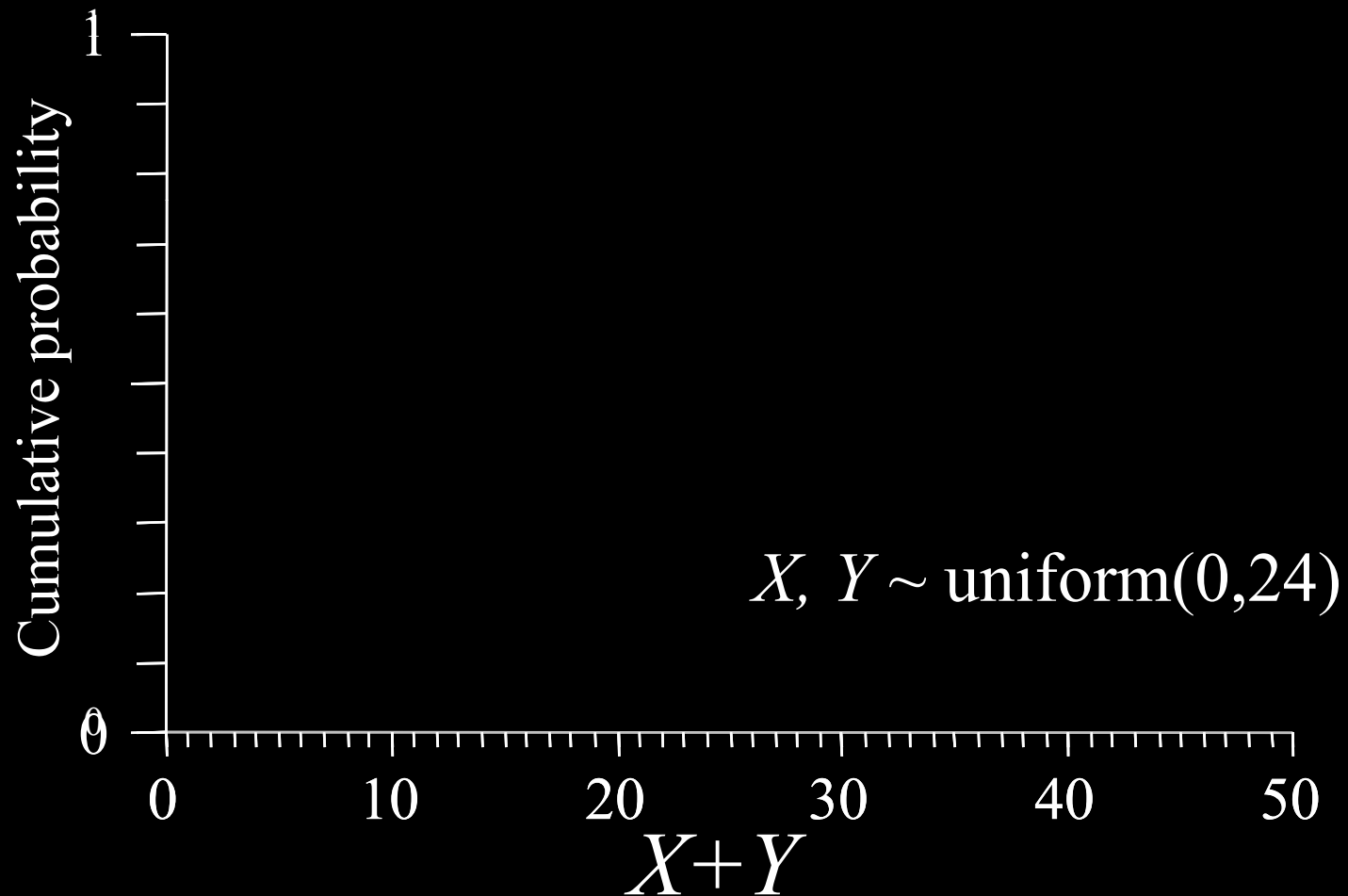
Perfect and opposite dependencies



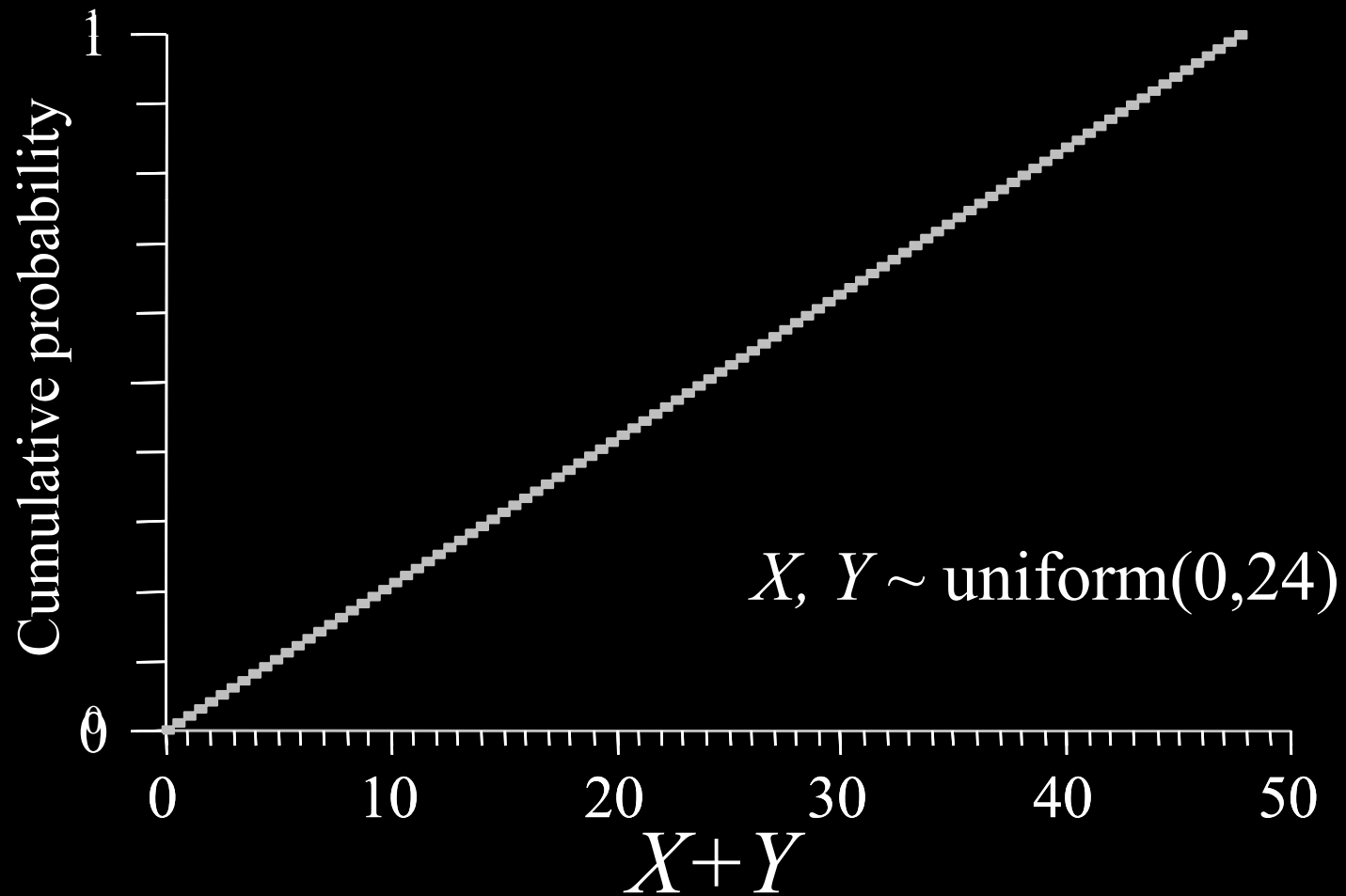
Uncertainty about dependence

- Sensitivity analyses usually used
 - Vary correlation coefficient between -1 and $+1$
- But this *underestimates* the true uncertainty
 - Example: suppose $X, Y \sim \text{uniform}(0,24)$ but we don't know the dependence between X and Y

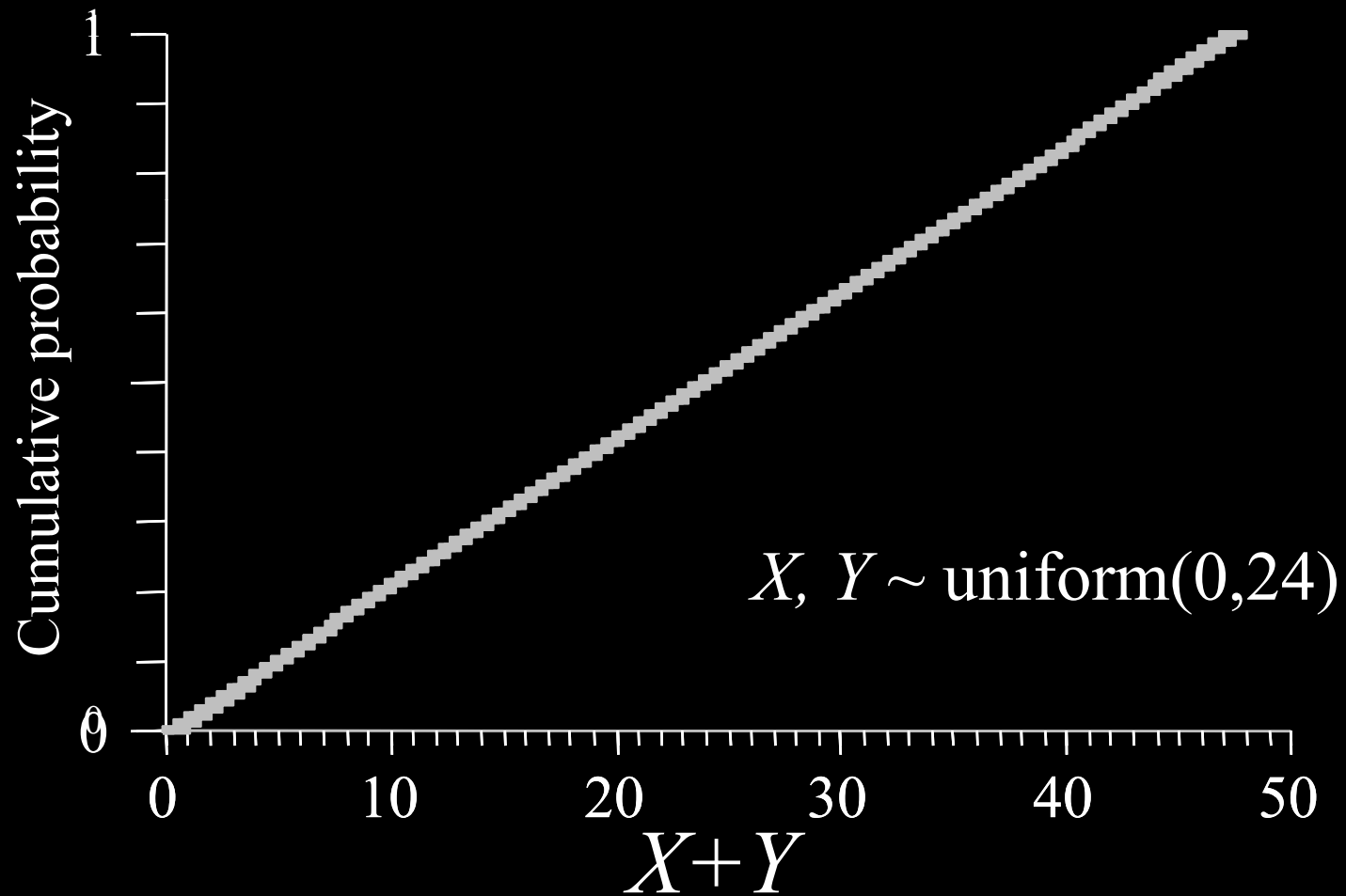
Varying the correlation coefficient



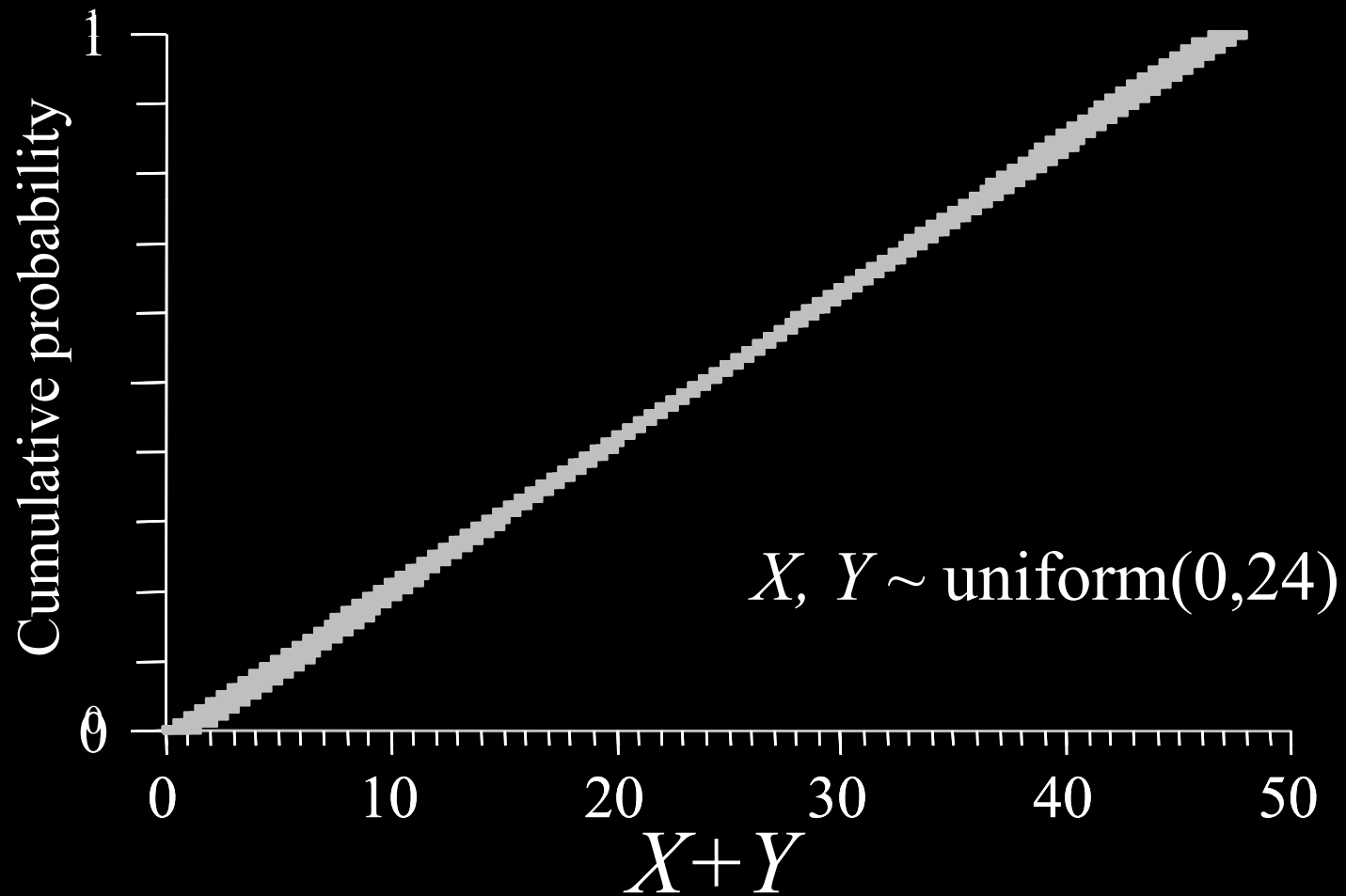
Varying the correlation coefficient



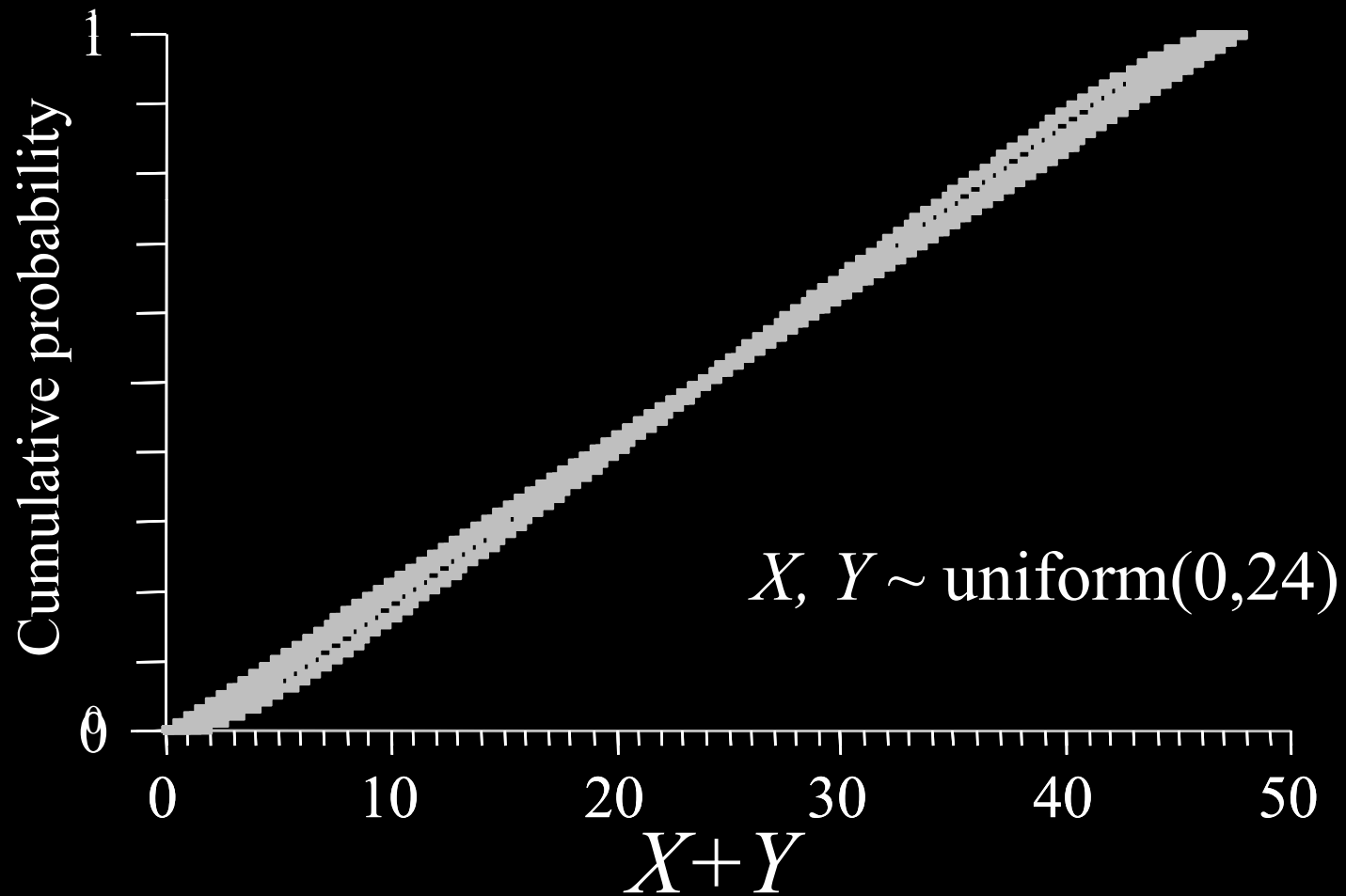
Varying the correlation coefficient



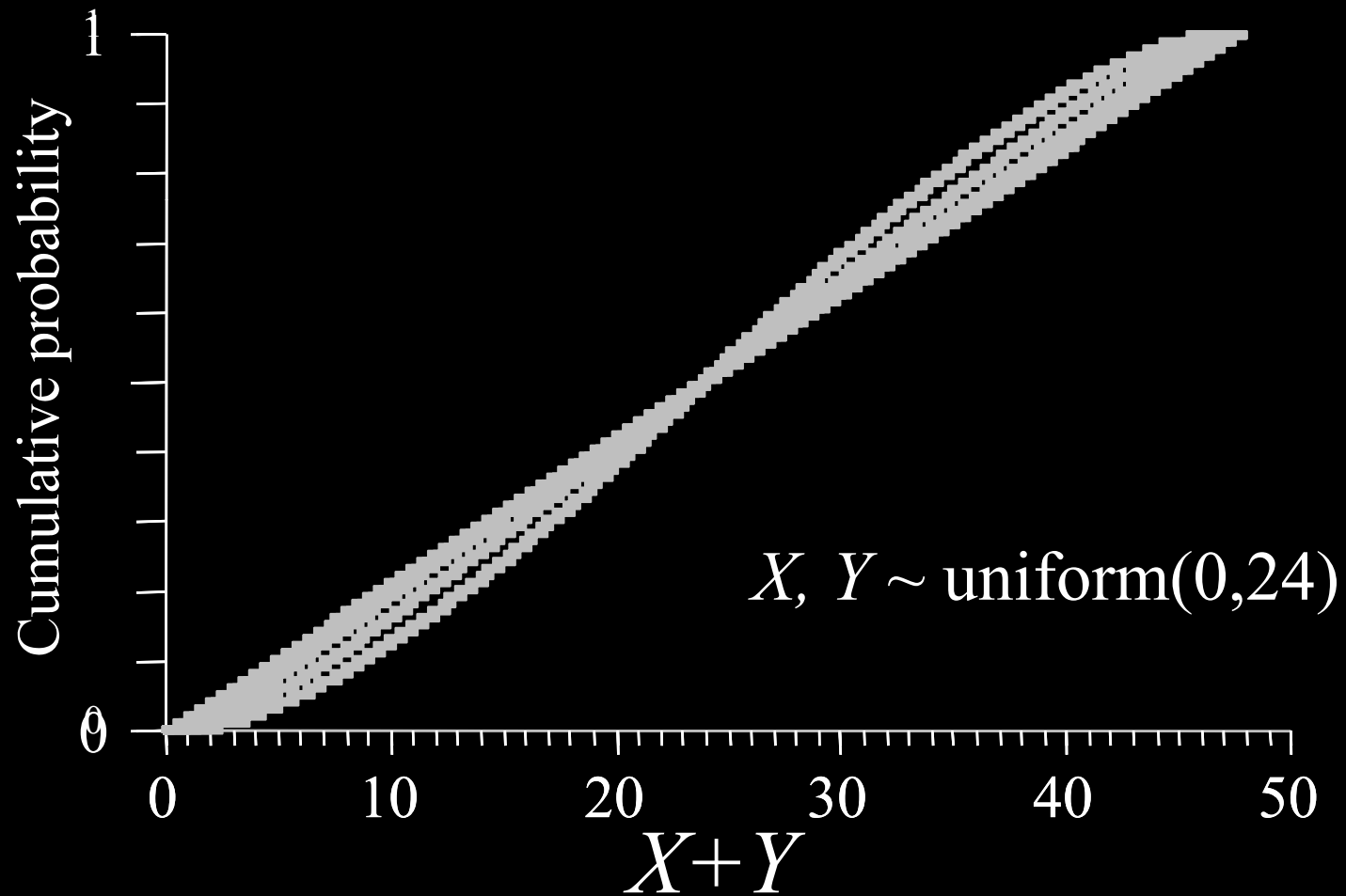
Varying the correlation coefficient



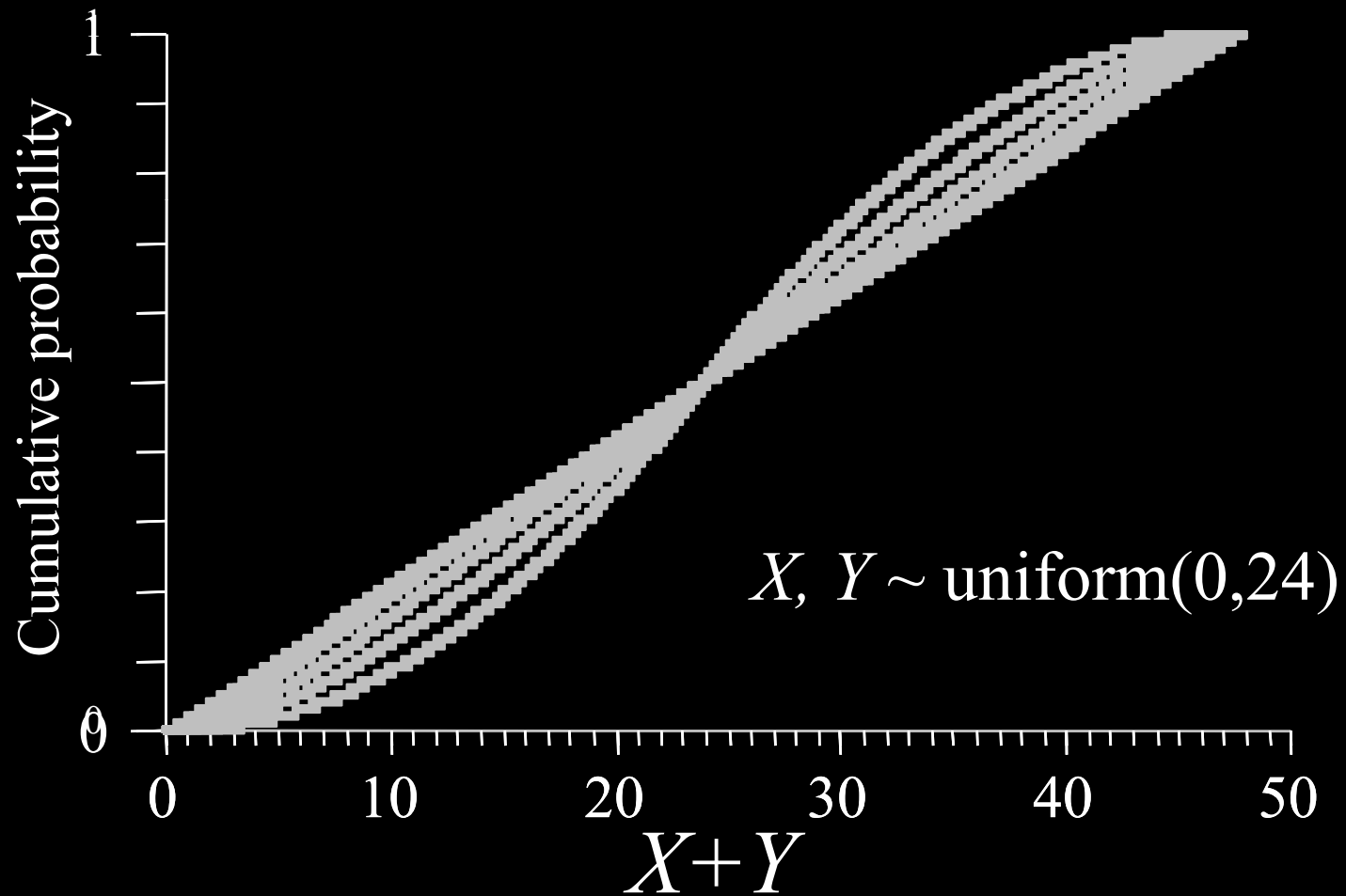
Varying the correlation coefficient



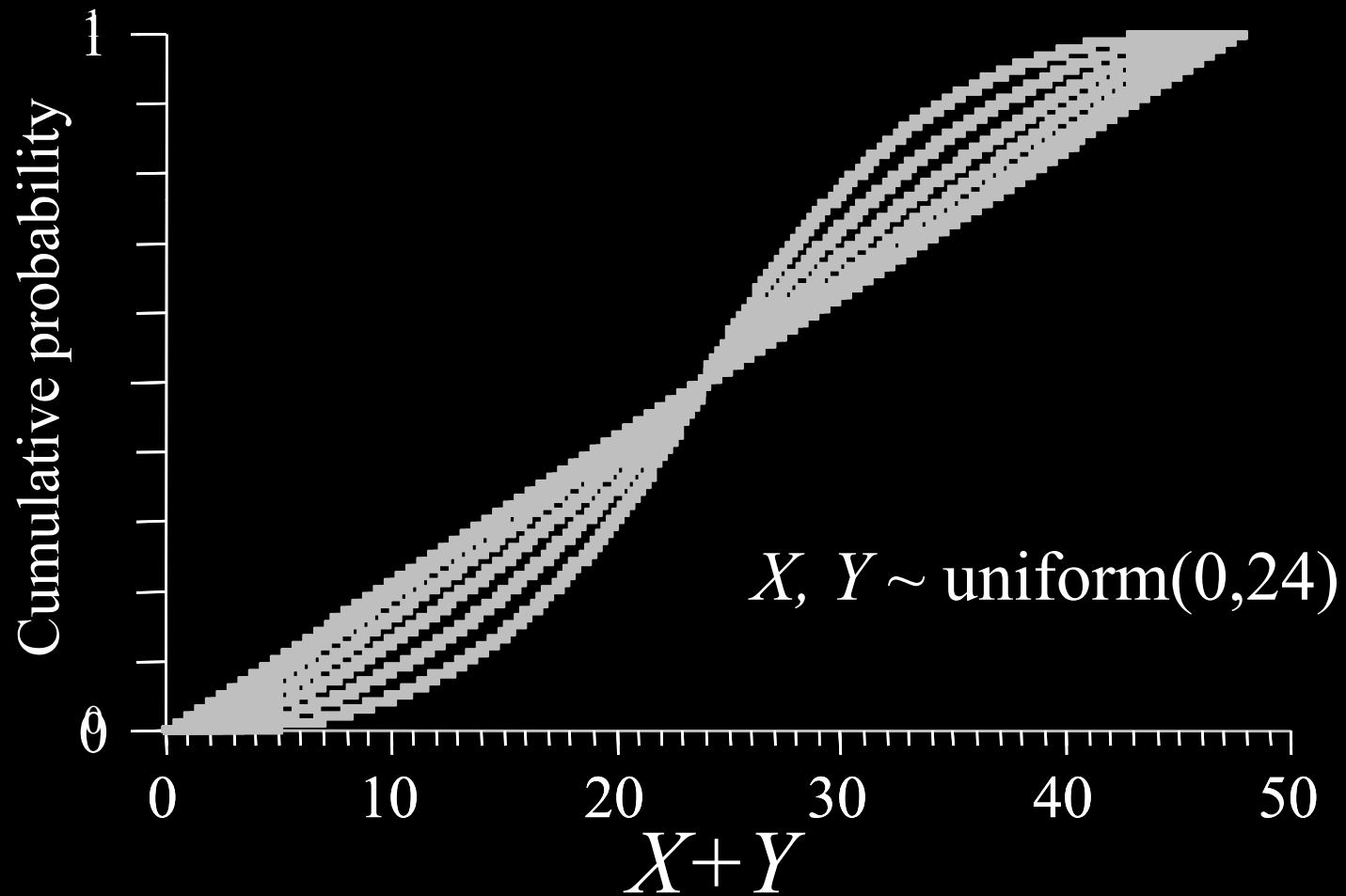
Varying the correlation coefficient



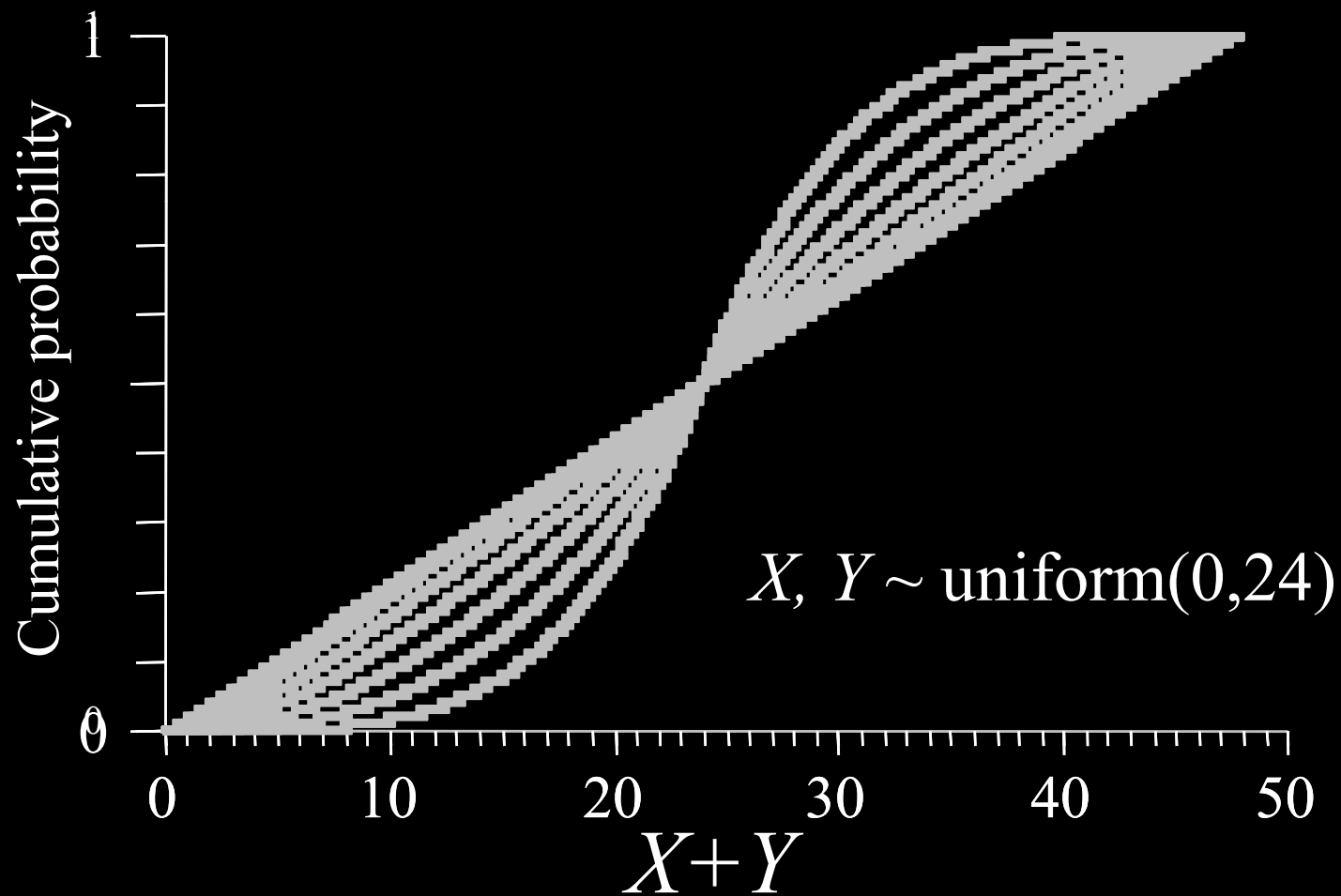
Varying the correlation coefficient



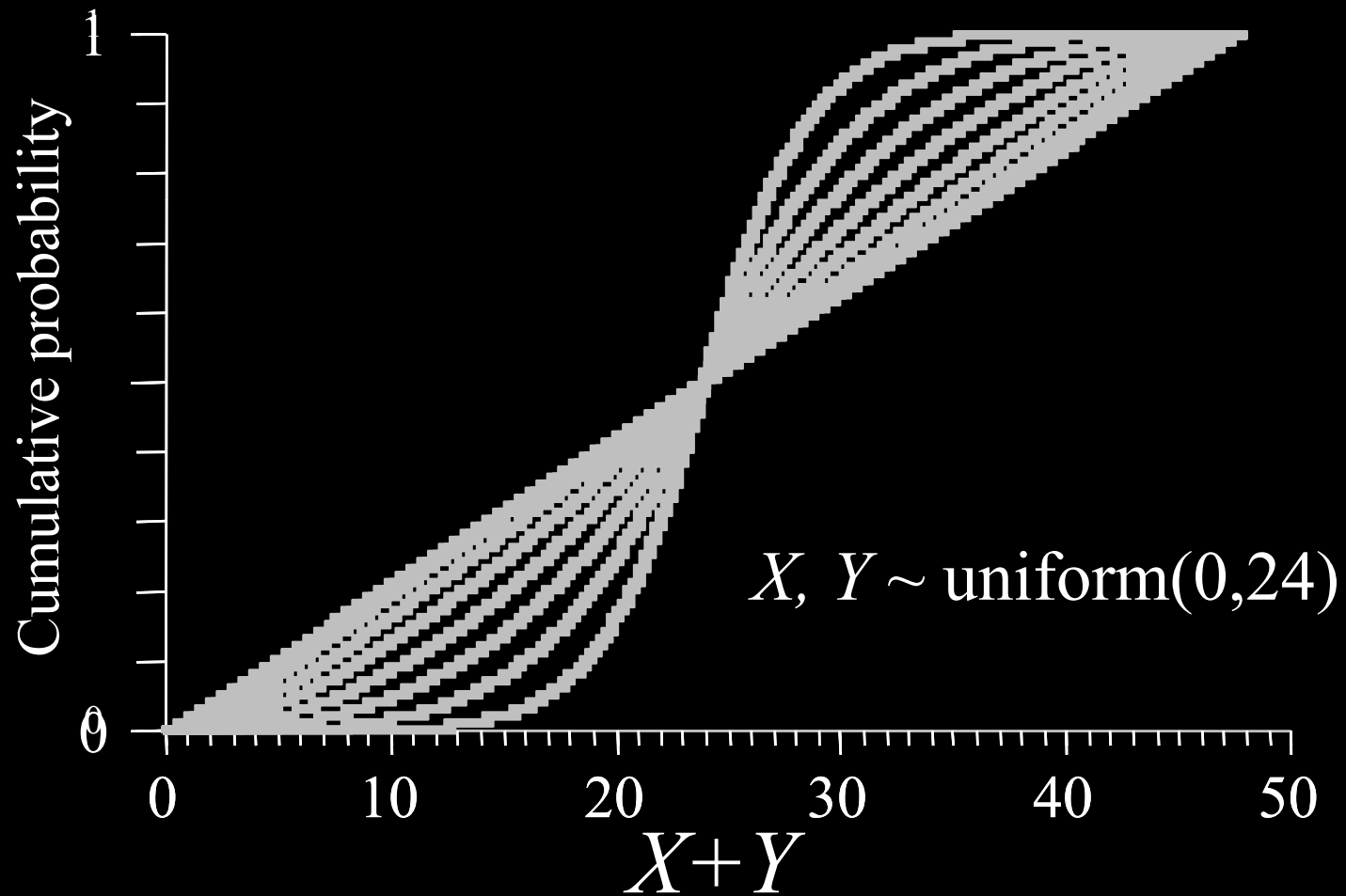
Varying the correlation coefficient



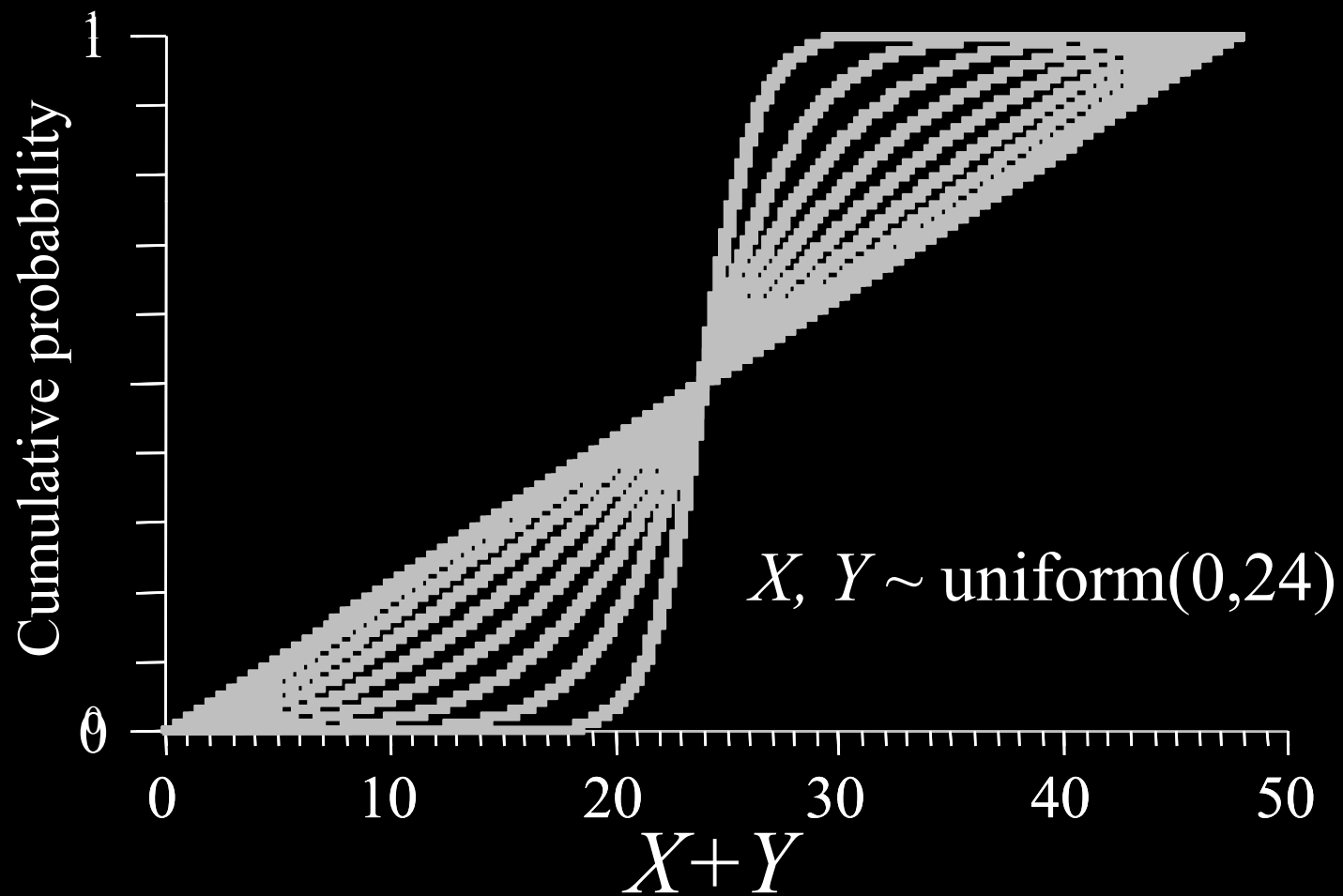
Varying the correlation coefficient



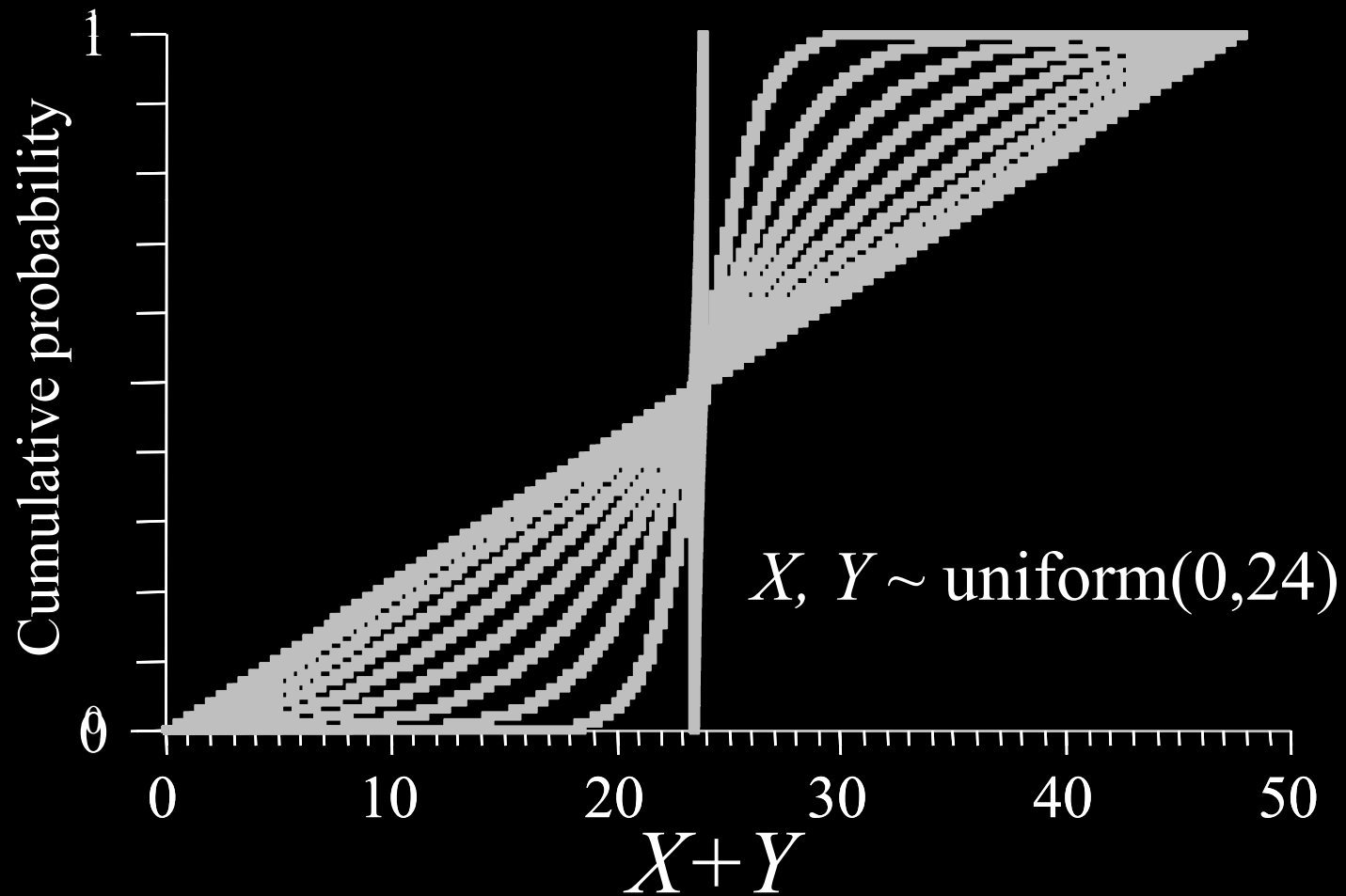
Varying the correlation coefficient



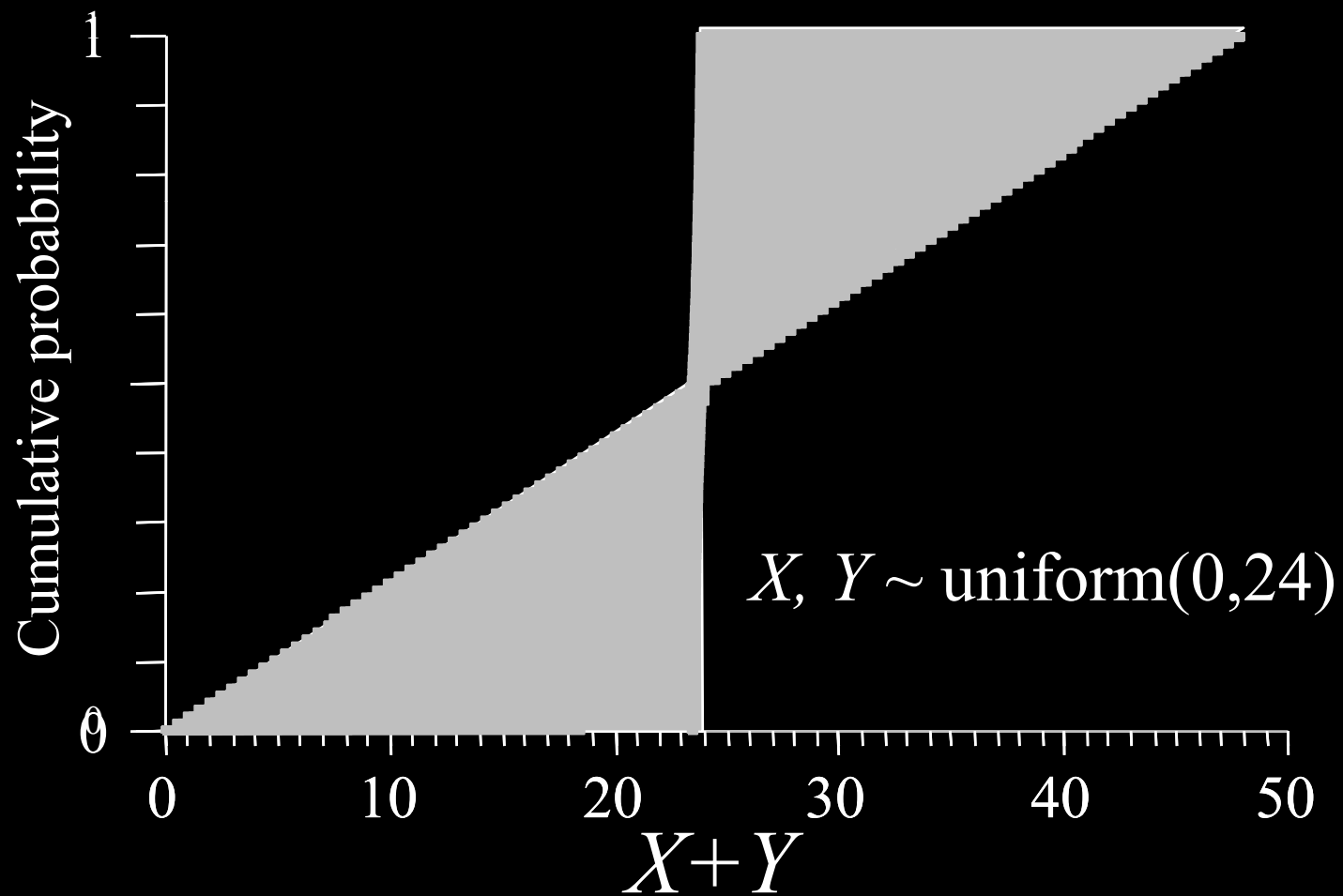
Varying the correlation coefficient



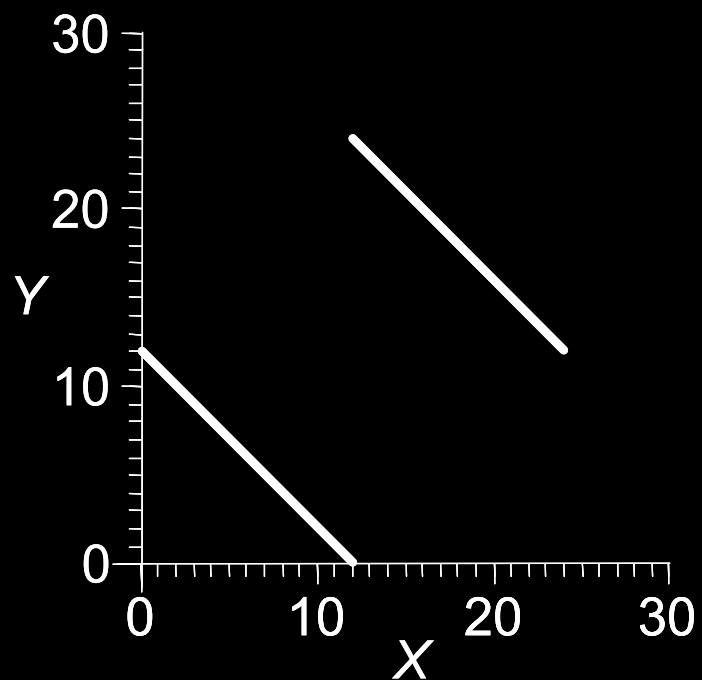
Varying the correlation coefficient



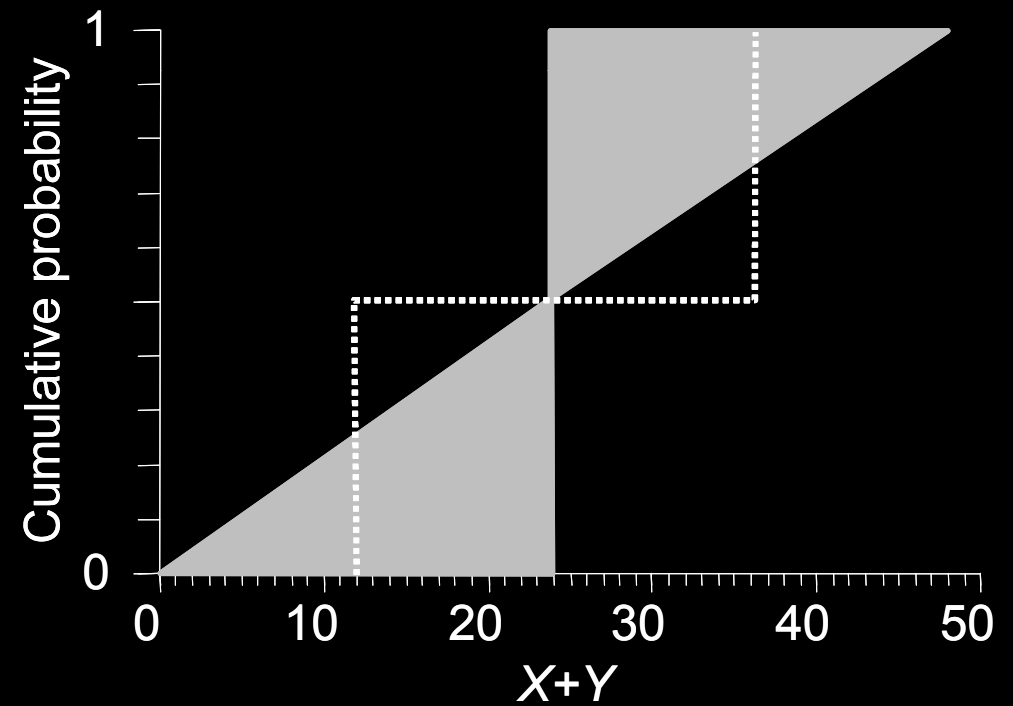
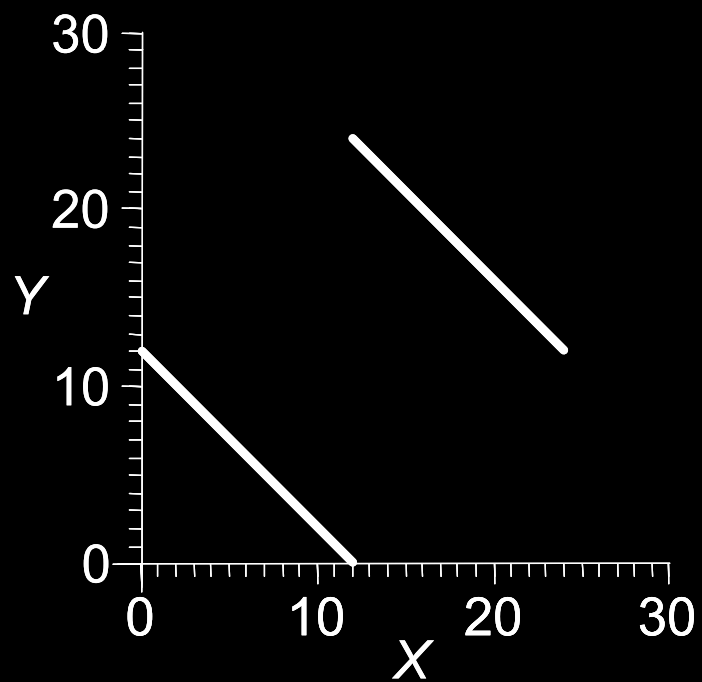
Varying the correlation coefficient



Counterexample



Counterexample



What about other dependencies?

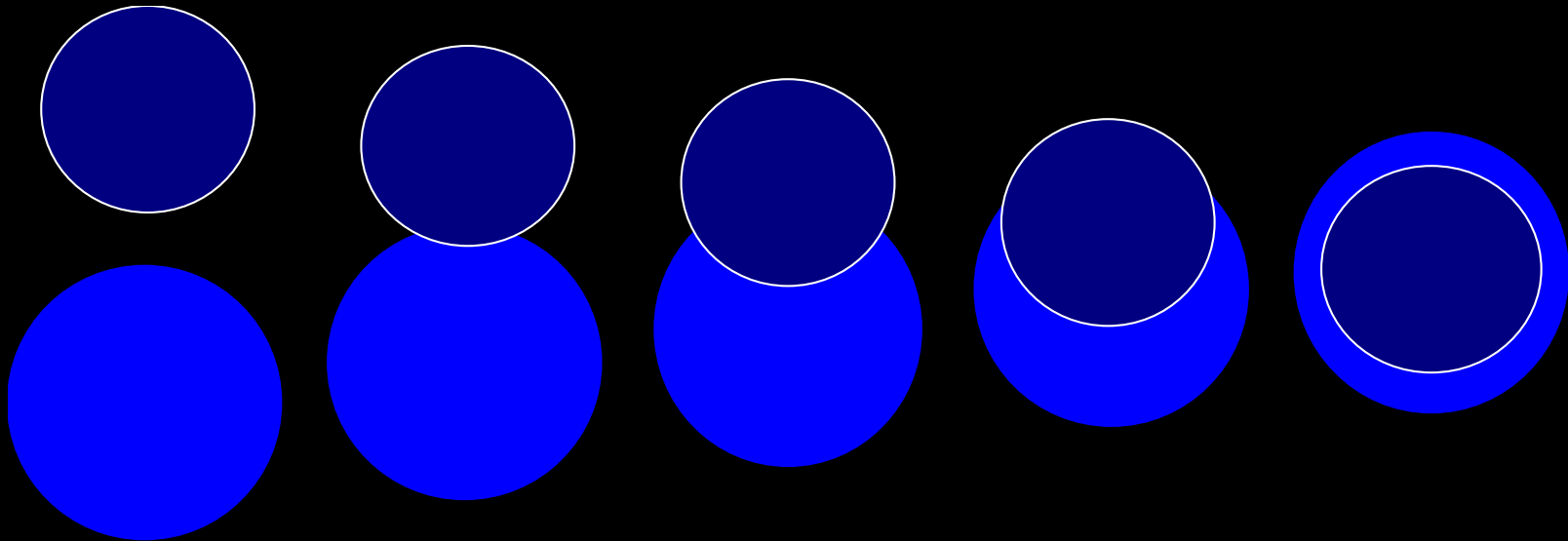
- Independent
- Perfectly positive
- Opposite
- Positively or negatively associated
- Specified correlation coefficient
- Nonlinear dependence (copula)
- Unknown dependence

Fréchet inequalities

They make *no assumption* about dependence (Fréchet 1935)

$$\max(0, P(A)+P(B)-1) \leq P(A \ \& \ B) \leq \min(P(A), P(B))$$

$$\max(P(A), P(B)) \leq P(A \vee B) \leq \min(1, P(A)+P(B))$$



Fréchet case (no assumption)

$A+B$

Fréchet case

$A \in [1,3]$
 $p_1 = 1/3$

$A \in [2,4]$
 $p_2 = 1/3$

$A \in [3,5]$
 $p_3 = 1/3$

$B \in [2,8]$
 $q_1 = 1/3$

$A+B \in [3,11]$
prob=[0,1/3]

$A+B \in [4,12]$
prob=[0,1/3]

$A+B \in [5,13]$
prob=[0,1/3]

$B \in [6,10]$
 $q_2 = 1/3$

$A+B \in [7,13]$
prob=[0,1/3]

$A+B \in [8,14]$
prob=[0,1/3]

$A+B \in [9,15]$
prob=[0,1/3]

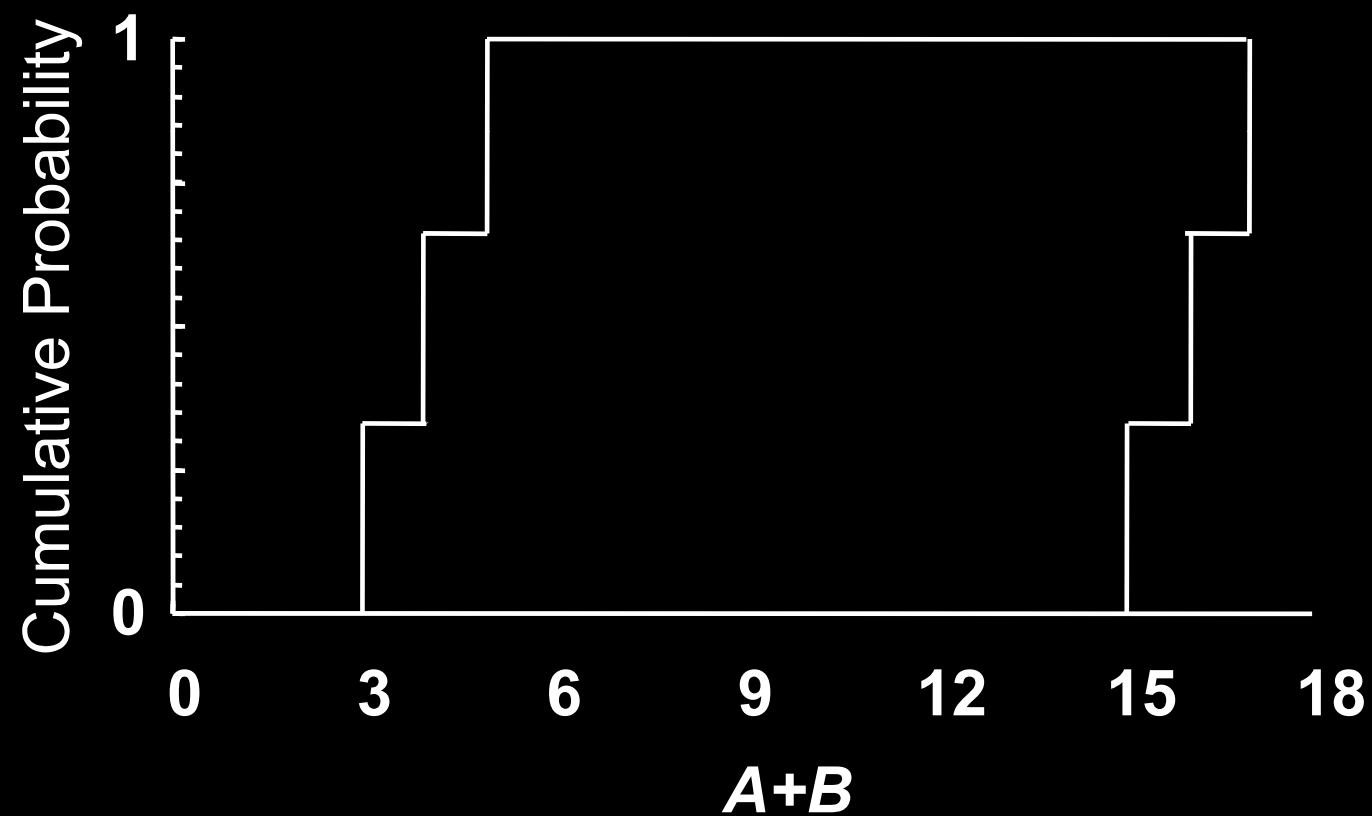
$B \in [8,12]$
 $q_3 = 1/3$

$A+B \in [9,15]$
prob=[0,1/3]

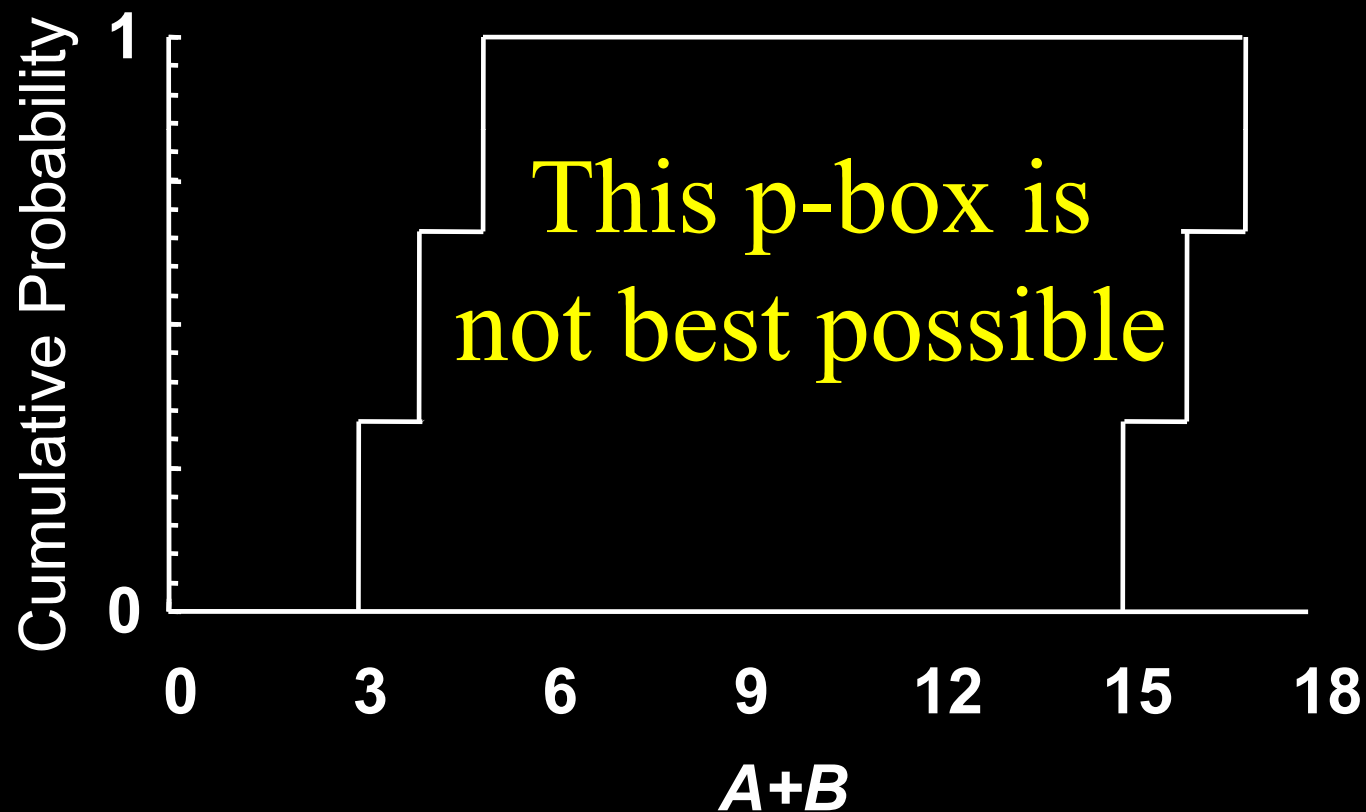
$A+B \in [10,16]$
prob=[0,1/3]

$A+B \in [11,17]$
prob=[0,1/3]

Naïve Fréchet case



Naïve Fréchet case



Fréchet can be improved

- Interval estimates of probabilities don't reflect the fact that the sum must equal one
- Resulting p-box is too fat
- Linear programming needed to get the optimal answer using this approach
- Frank, Nelsen and Sklar gave a way to compute the optimal answer directly

Frank, Nelsen and Sklar (1987)

If $X \sim F$ and $Y \sim G$, then the distribution of $X+Y$ is

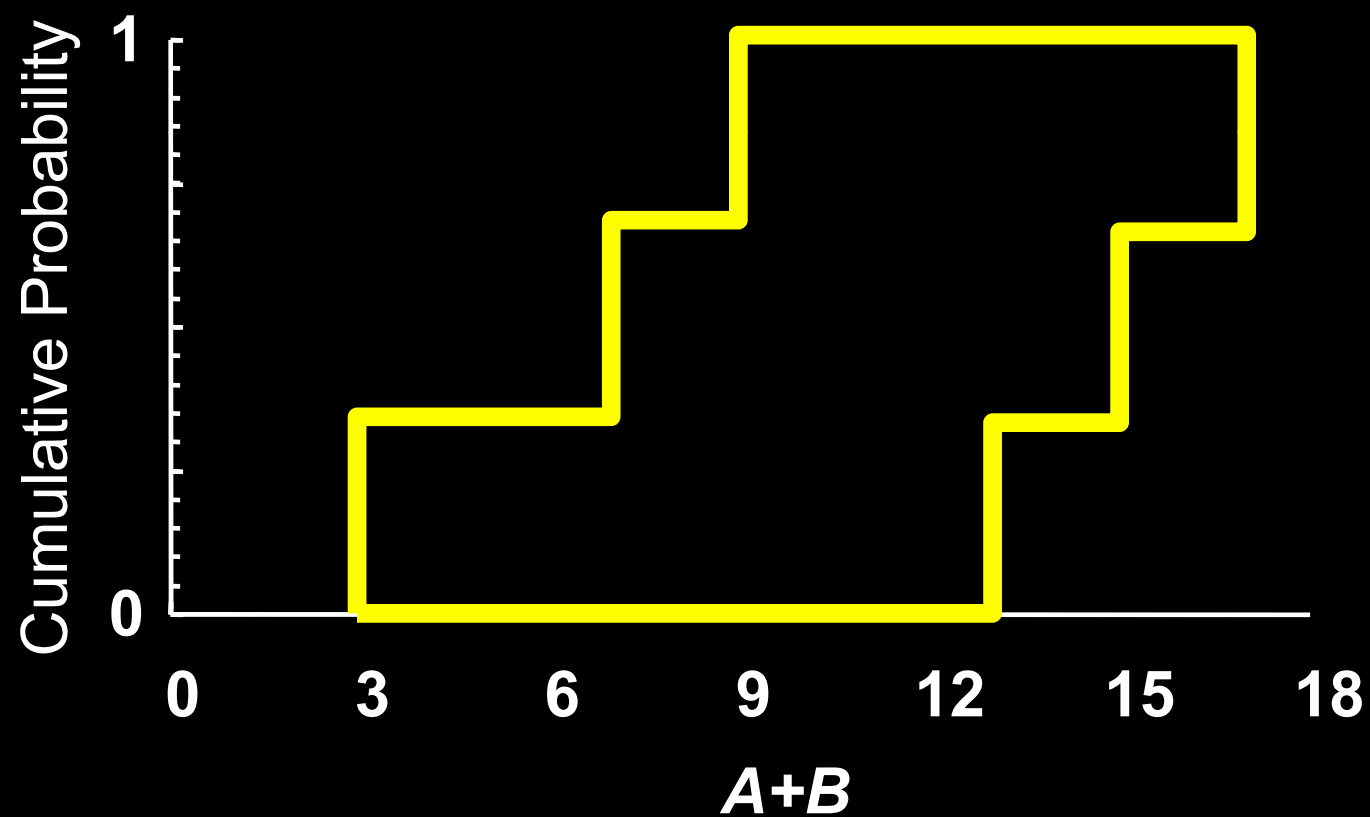
$$\sigma_{+,C}(F,G)(z) = \int_{x+y < z} dC(F(x), G(y))$$

where C is the copula between F and G . In any case, and irrespective of this dependence, this distribution is bounded by

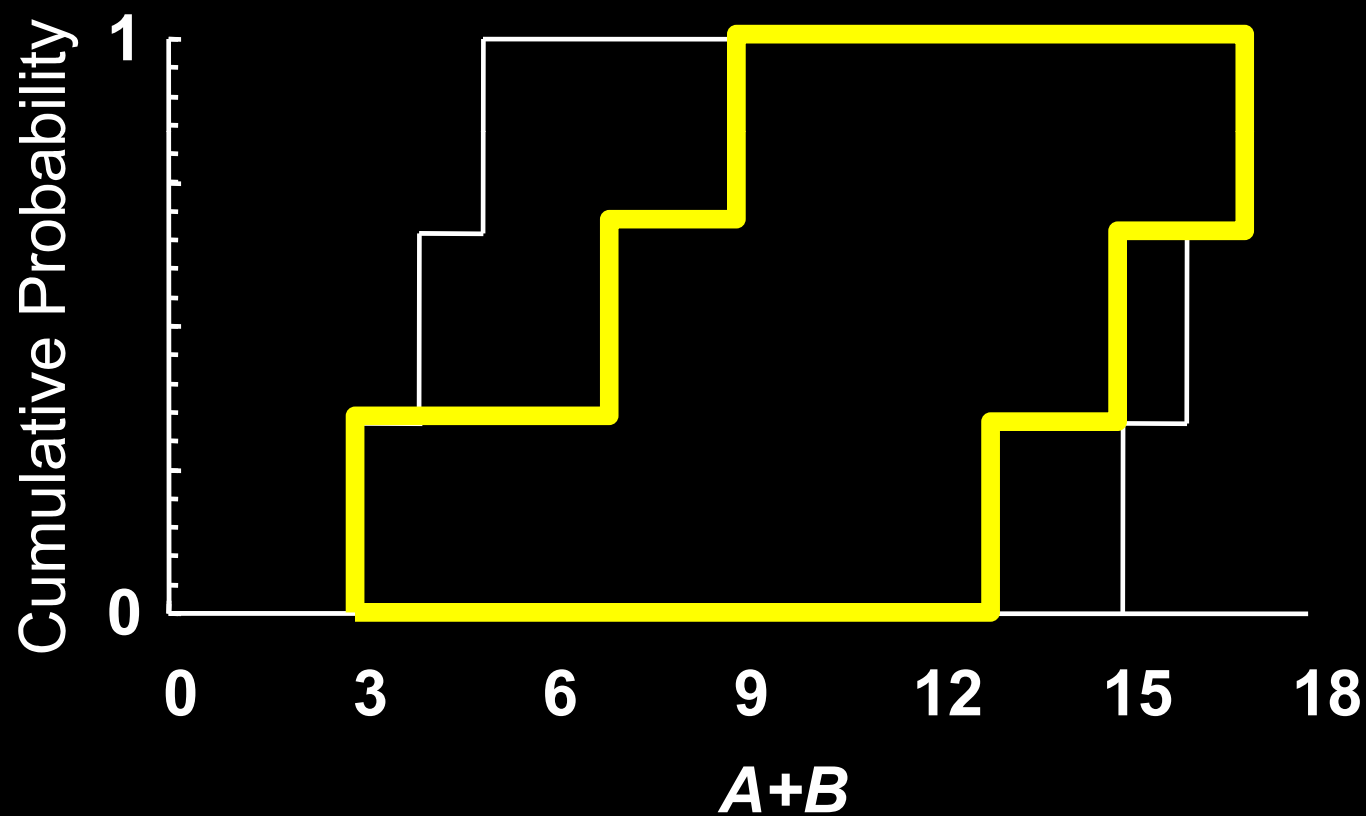
$$\left[\sup_{z=x+y} \max(F(x) + G(y) - 1, 0), \inf_{z=x+y} \min(F(x) + G(y), 1) \right]$$

This formula can be generalized to work with bounds on F and G .

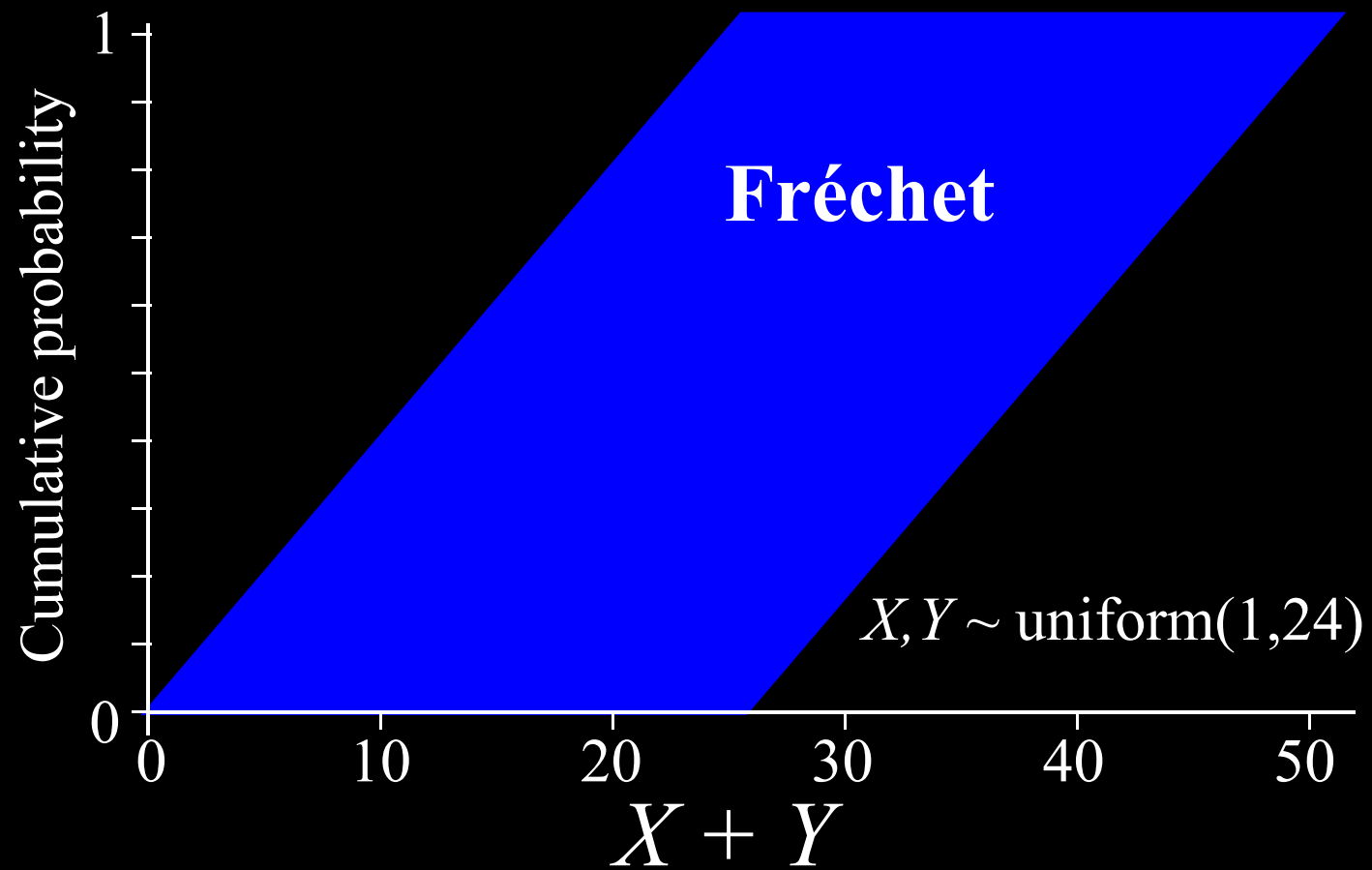
Best possible bounds



Best possible bounds



Unknown dependence



Between independence and Fréchet

- Some information may be available by which the p-boxes could be tightened over the Fréchet case without specifying the dependence perfectly, e.g.,

- Dependence is positive (PQD)

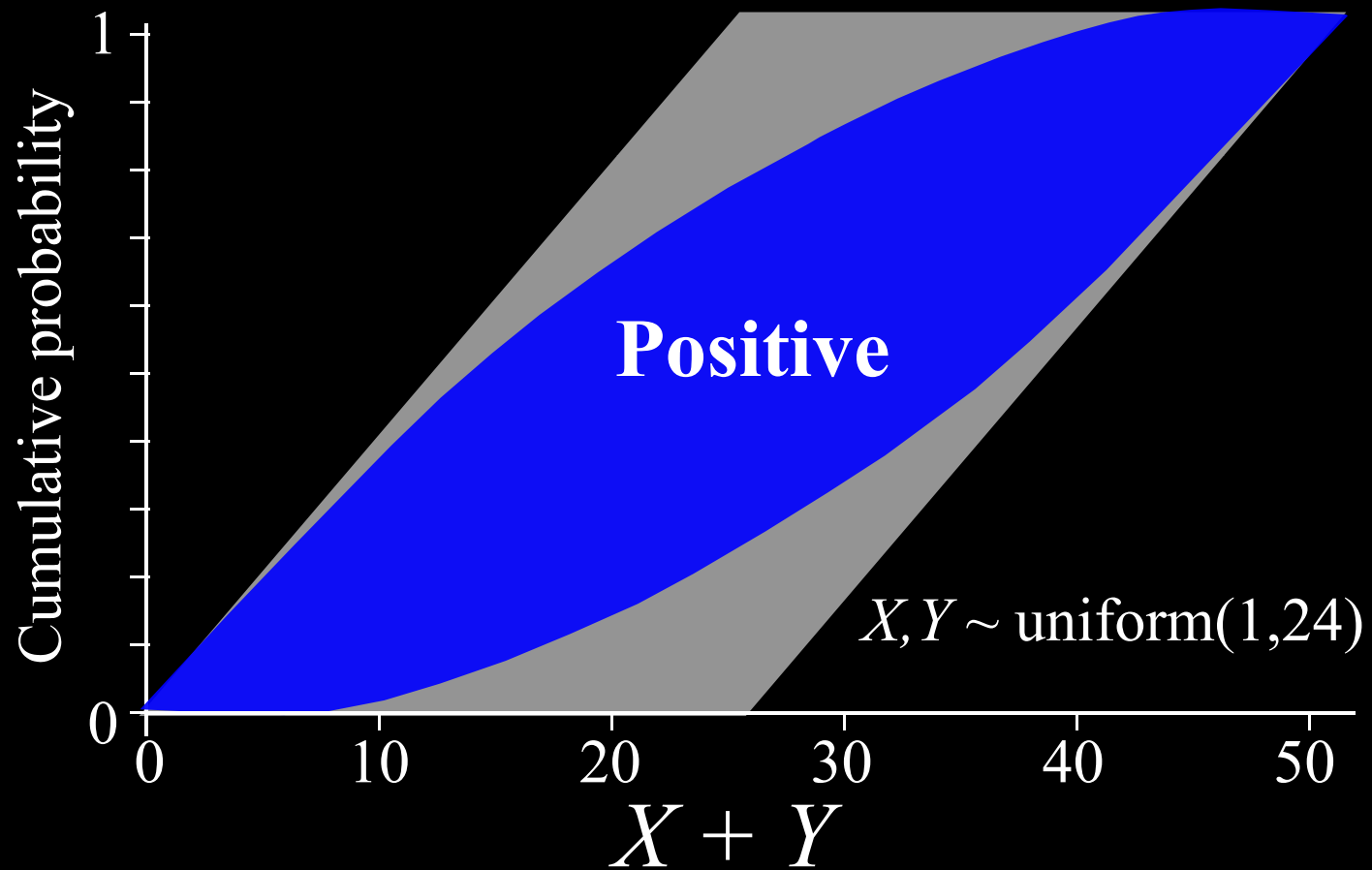
$P(X \leq x, Y \leq y) \geq P(X \leq x) P(Y \leq y)$ for all x and y

$$\left[\sup_{z=x+y} (F(x)G(y)), \inf_{z=x+y} (1 - (1 - F(x))(1 - G(y))) \right]$$

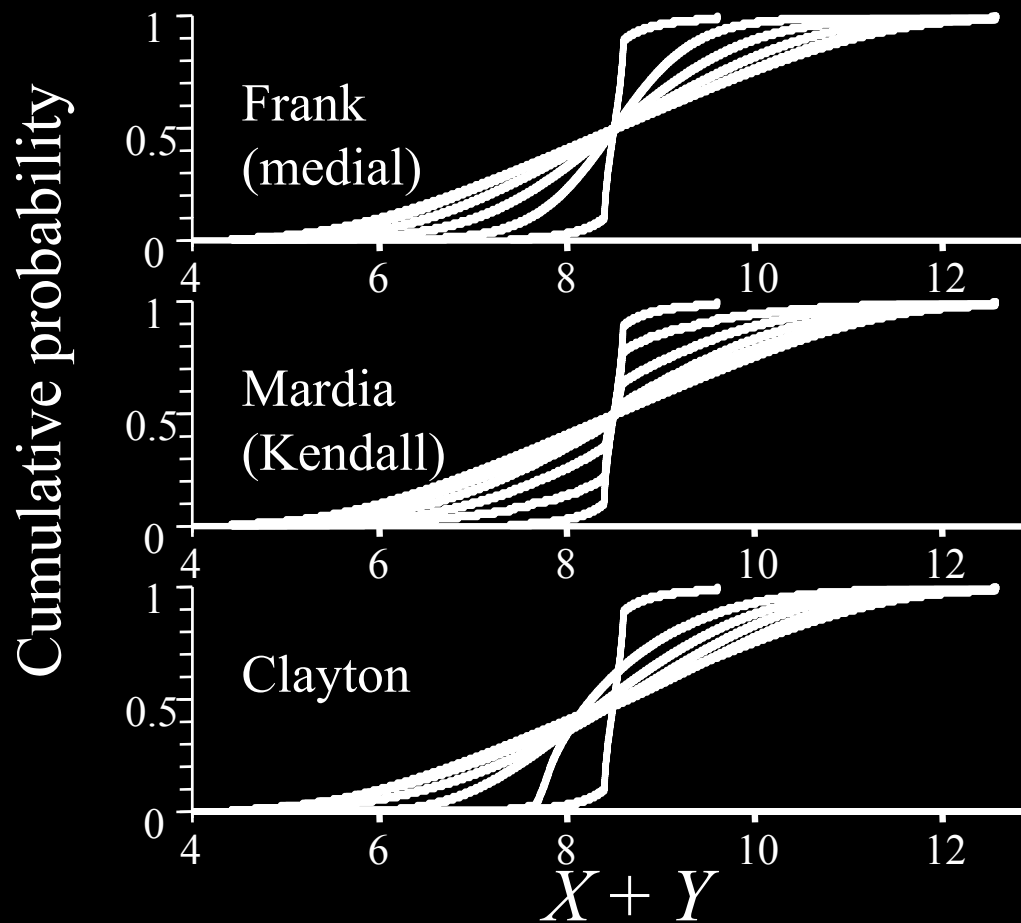
- Variables are uncorrelated

Pearson correlation r is zero

Unknown but positive dependence

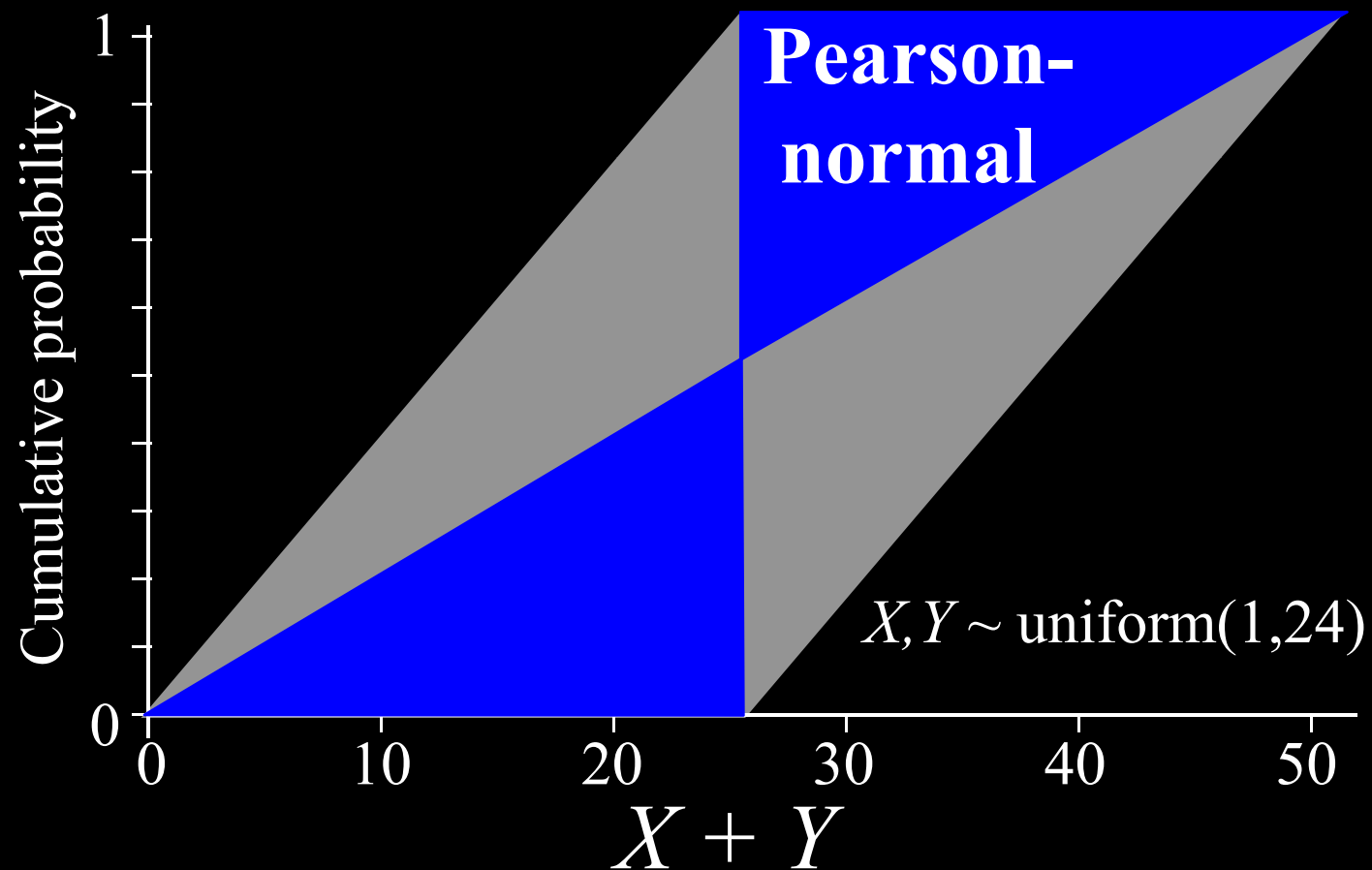


Can model dependence exactly too

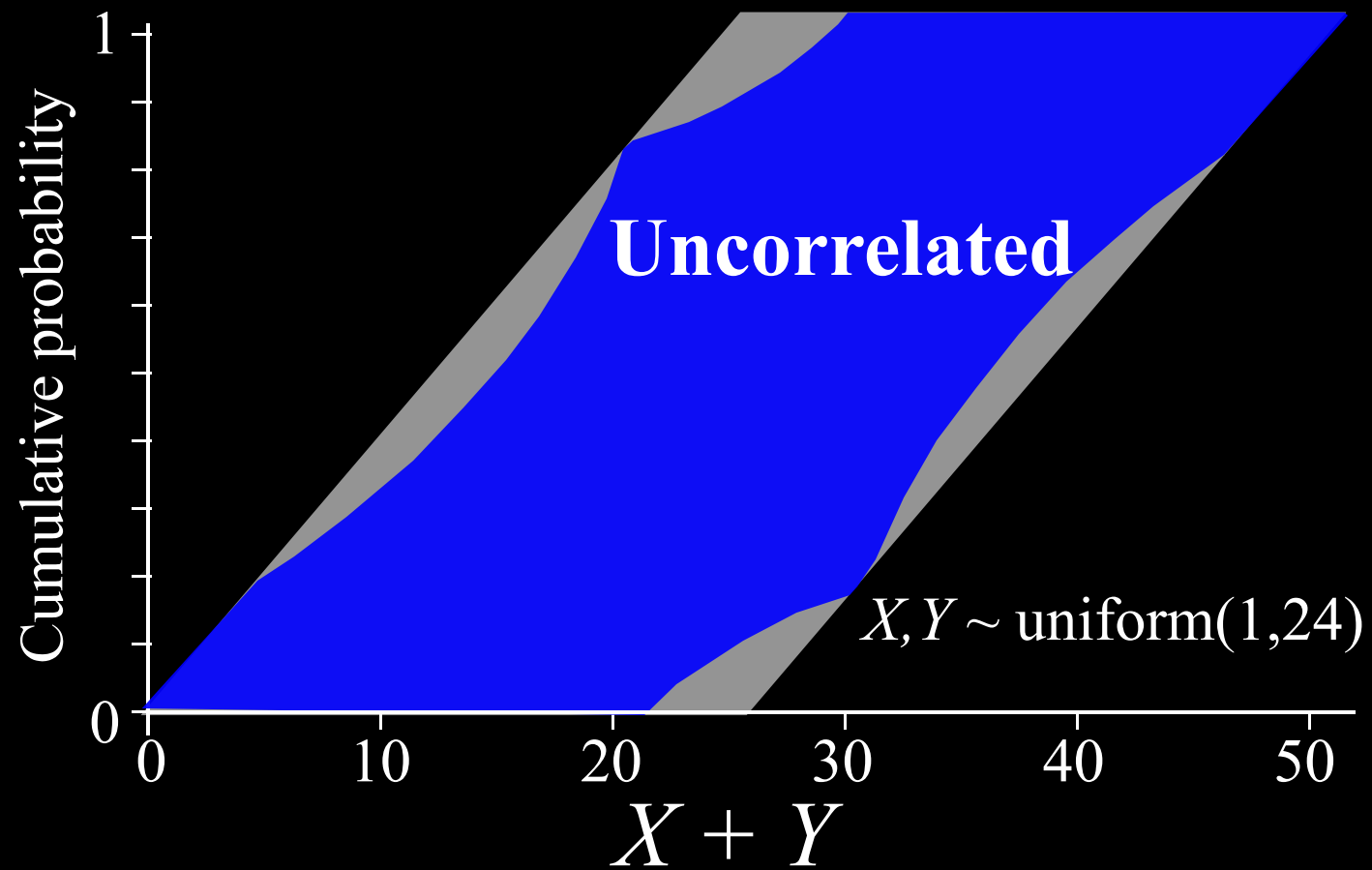


$X \sim \text{normal}(5,1)$
 $Y \sim \text{uniform}(2,5)$
various correlations
and dependence
functions (copulas)

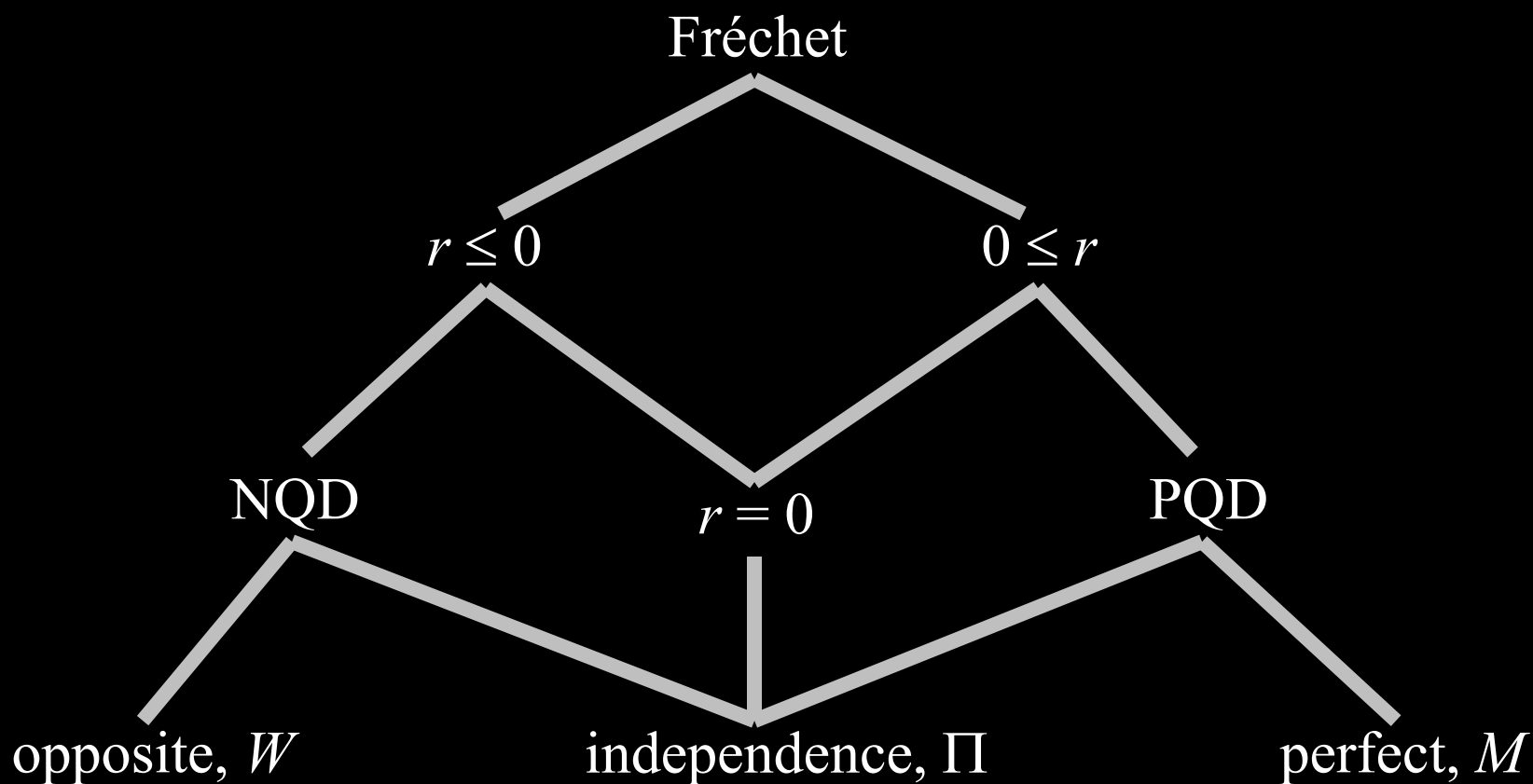
Varying correlation between -1 and $+1$



Uncorrelated variables




Dependence tree




Dependence between p-boxes

- Random-set independent
 - Epistemically independent
 - Strongly independent
 - Repetition independent
- Perfectly associated
- Oppositely associated
- Known copula
- Specified copula family and correlation
- Known functional relationship
- Positively quadrant dependent (PQD)
- Negatively quadrant dependent (NQD)
- Known or interval-bounded correlation
- Fréchet case

For precise probabilities, these are all the same.



These cases yield precise distributions from precise input distributions



Example: dioxin inhalation

Location: Superfund site in California

Receptor: adults in neighboring community

Contaminant: dioxin

Exposure route: inhalation of windborne soil

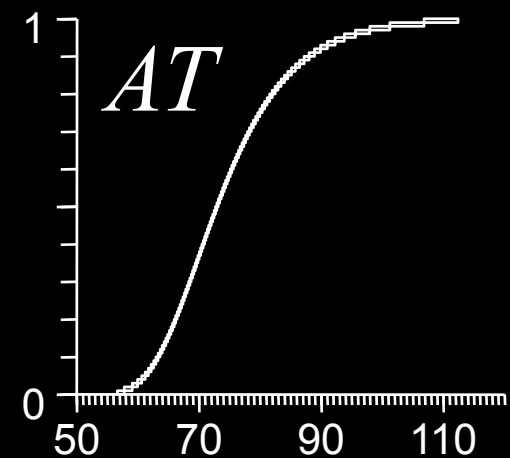
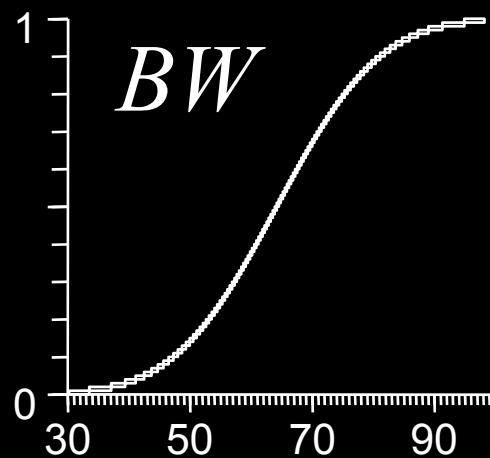
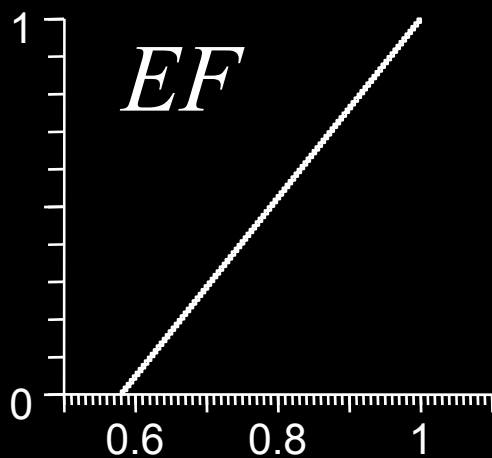
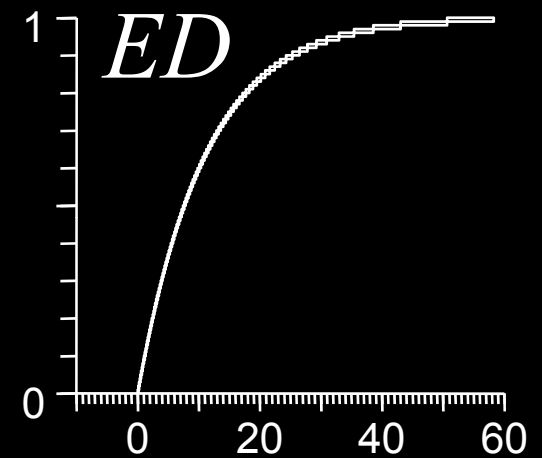
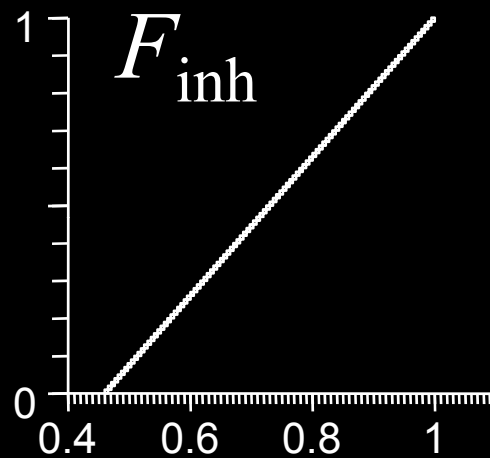
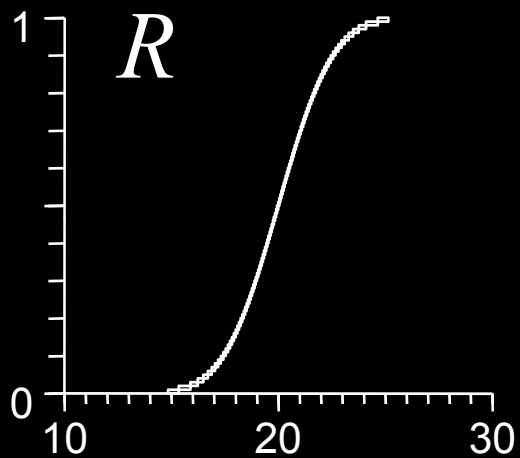
Modified from Table II and IV in Copeland, T.L., A.M. Holbrow, J.M. Otani, K.T. Conner and D.J. Paustenbach 1994. Use of probabilistic methods to understand the conservatism in California's approach to assessing health risks posed by air contaminant. *Journal of the Air and Waste Management Association* 44: 1399-1413.

Total daily intake from inhalation

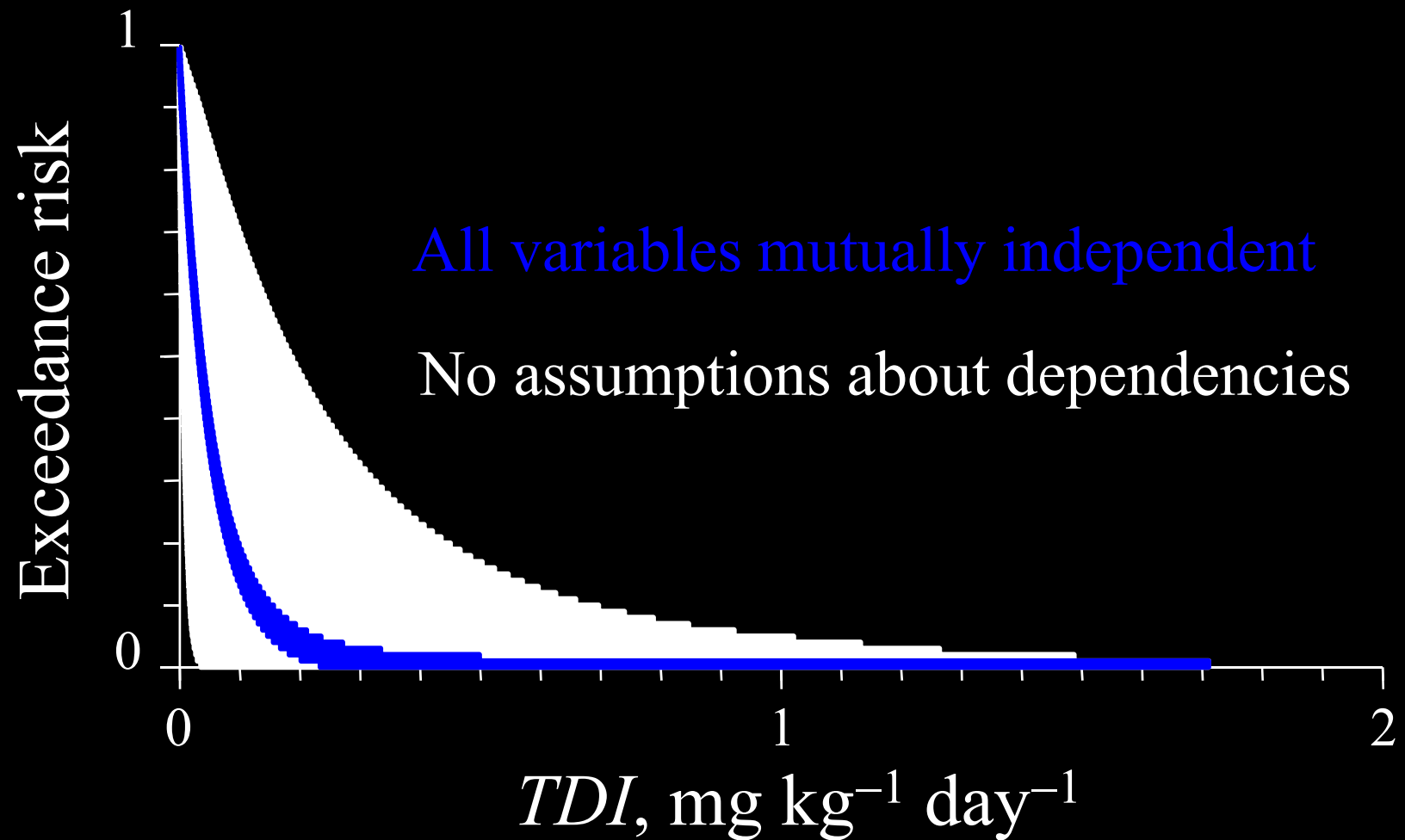
$$T_{DI} = \frac{R \cdot C_{GL} \cdot F_{inh} \cdot ED \cdot EF}{BW \cdot AT}$$

$R = \text{normal}(20, 2)$ //respiration rate, m³/day
 $C_{GL} = 2$ //concentration at ground level, mg/m³
 $F_{inh} = \text{uniform}(0.46, 1)$ //fraction of particulates retained in lung, [unitless]
 $ED = \text{exponential}(11)$ //exposure duration, years
 $EF = \text{uniform}(0.58, 1)$ //exposure frequency, fraction of a year
 $BW = \text{normal}(64.2, 13.19)$ //receptor body weight, kg
 $AT = \text{gumbel}(70, 8)$ //averaging time, years

Input distributions



Results



Uncertainty about dependence

- Impossible with sensitivity analysis since it's an infinite-dimensional problem
- Kolmogorov-Fréchet bounding lets you be sure
- Can be a large or a small consequence

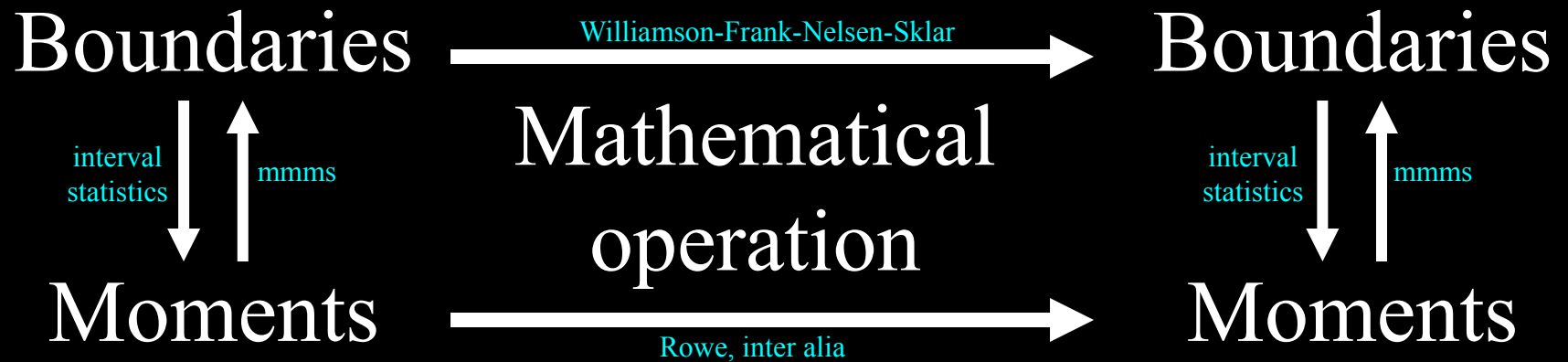
Further wrinkles

Simultaneous moment propagation

- Just means and variances, and ranges
- Makes use of general formulas, and also special formulas for named distribution shapes
- Finite ranges imply moments always exist and often improve bounds formulas substantially
- Intersects bounds from formulas and inferred from distribution bounds

Inputs

Output



This auxiliary effort often *substantially* improves (tightens) the output p-boxes

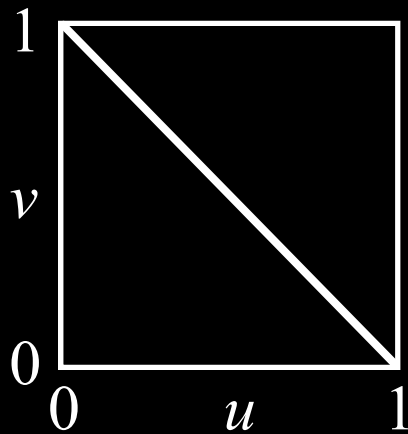
Interval dependence *within* cells

- Interval arithmetic traditionally does not address the dependence between its arguments
- But if we extend it so it does, we can improve the results

Three special cases

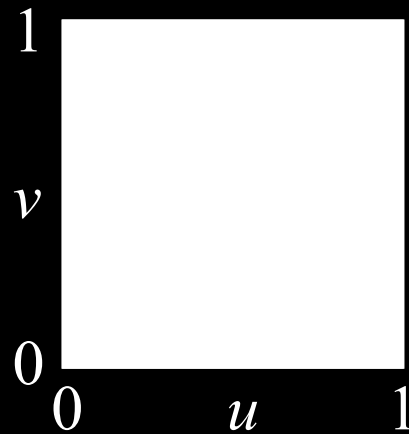
Opposite

(countermonotonic)



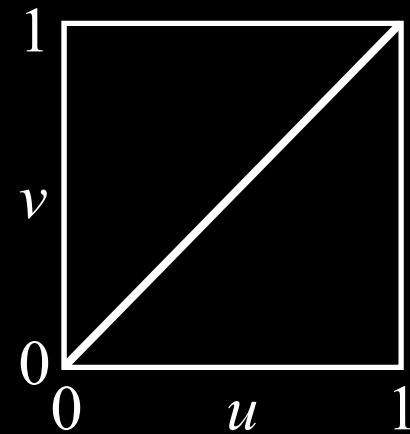
Nondependent

(the Fréchet case)



Perfect

(comonotonic)



Correlation

- A model of dependence that's parameterized by a (scalar) value called the “correlation coefficient”

$$\rho : [-1, +1] \rightarrow \mathcal{D} = 2^{[0,1] \times [0,1]}$$

- A correlation model is called “complete” if

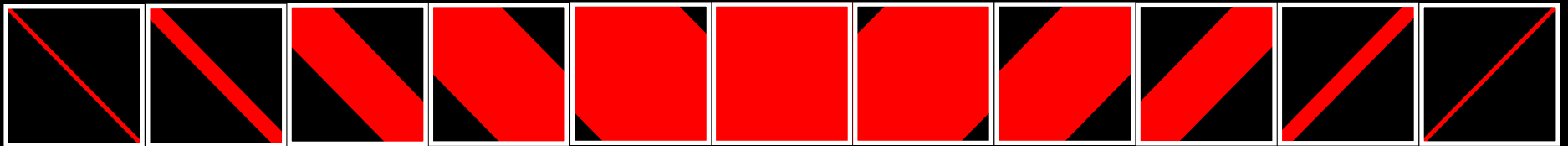
$$\rho(-1) = \square_{\diagdown}, \quad \rho(0) = \blacksquare, \quad \rho(+1) = \square_{\diagup}$$

Corner-shaving dependence

$r = -1$

$r = 0$

$r = +1$



$$D(r) = \{ (u,v) : \max(0, -u-r, u-1+r) \leq v \leq \min(1, u+1-r, -u+2+r) \}$$

$$u \in [0,1], v \in [0,1]$$

$$f(A, B) = \{ c : c = f(u(a_2-a_1)+a_1, v(b_2-b_1)+b_1), (u,v) \in D \}$$

$$A+B = [\text{env}(w(A, -r)+b_1, a_1+w(B, -r)), \text{env}(a_2+w(B, 1+r), w(A, 1+r)+b_2)]$$

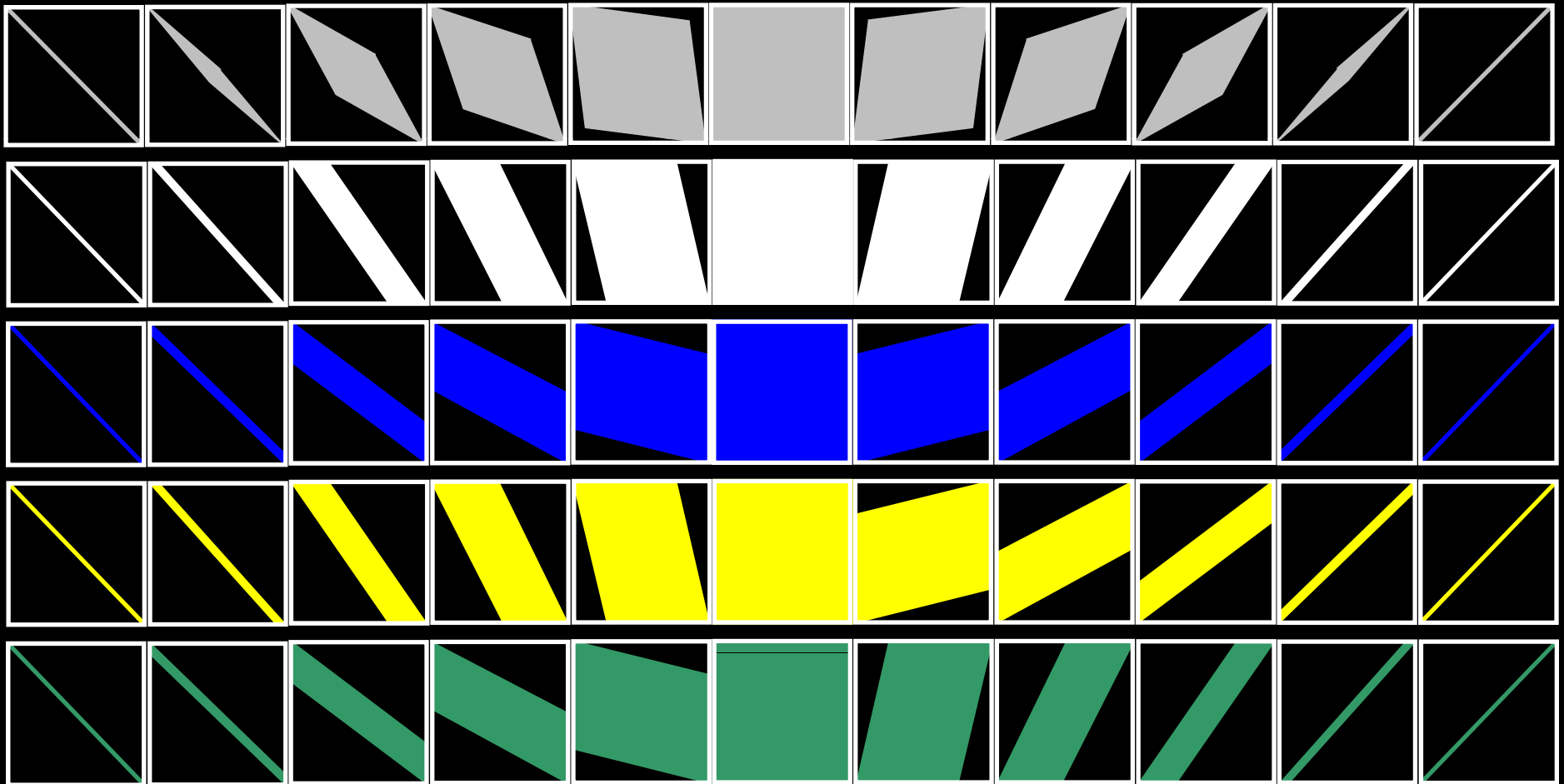
$$w([a_1, a_2], p) = \begin{cases} a_1 & \text{if } p < 0 \\ a_2 & \text{if } 1 < p \\ p(a_2-a_1)+a_1 & \text{otherwise} \end{cases}$$

Other complete correlation families

$r = -1$

$r = 0$

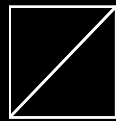
$r = +1$



$A + B$

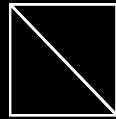
$A = [2,5]$

$B = [3,9]$



[5, 14]

Perfect



[8, 11]

Opposite



[7.1, 11.9]

Corner-shaving ($r = -0.7$)



[7.27, 11.73] Elliptic ($r = -0.7$)



[5, 14]

Upper, left



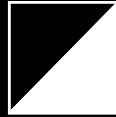
[5, 11]

Lower, left



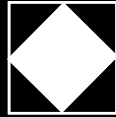
[8, 14]

Upper, right



[5, 14]

Lower, right



[6.5, 12.5]

Diamond



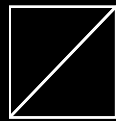
[5, 14]

Nondependent

$$A + B$$

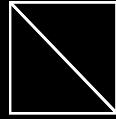
$$A = [2, 5]$$

$$B = [3, 9]$$



[5, 14]

Perfect



[8, 11]

Opposite



[7.1, 11.9]

Corner-shaving ($r = -0.7$)



[7.27, 11.73] Elliptical ($r = -0.7$)



[5, 14]

Upper, left



[5, 14]

Lower, left



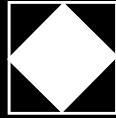
[8, 14]

Upper, right



[5, 14]

Lower, right



[6.5, 12.5]

Diamond



[5, 14]

Nondependent

Tighter!

Opposite/nondependent

$A+B$

opposite
nondependent

$A \in [1,3]$
 $p_1 = 1/3$

$A \in [2,4]$
 $p_2 = 1/3$

$A \in [3,5]$
 $p_3 = 1/3$

$B \in [2,8]$
 $q_1 = 1/3$

$A+B \in [3,11]$
prob=0

$A+B \in [4,12]$
prob=0

$A+B \in [5,13]$
prob=1/3

$B \in [6,10]$
 $q_2 = 1/3$

$A+B \in [7,13]$
prob=0

$A+B \in [8,14]$
prob=1/3

$A+B \in [9,15]$
prob=0

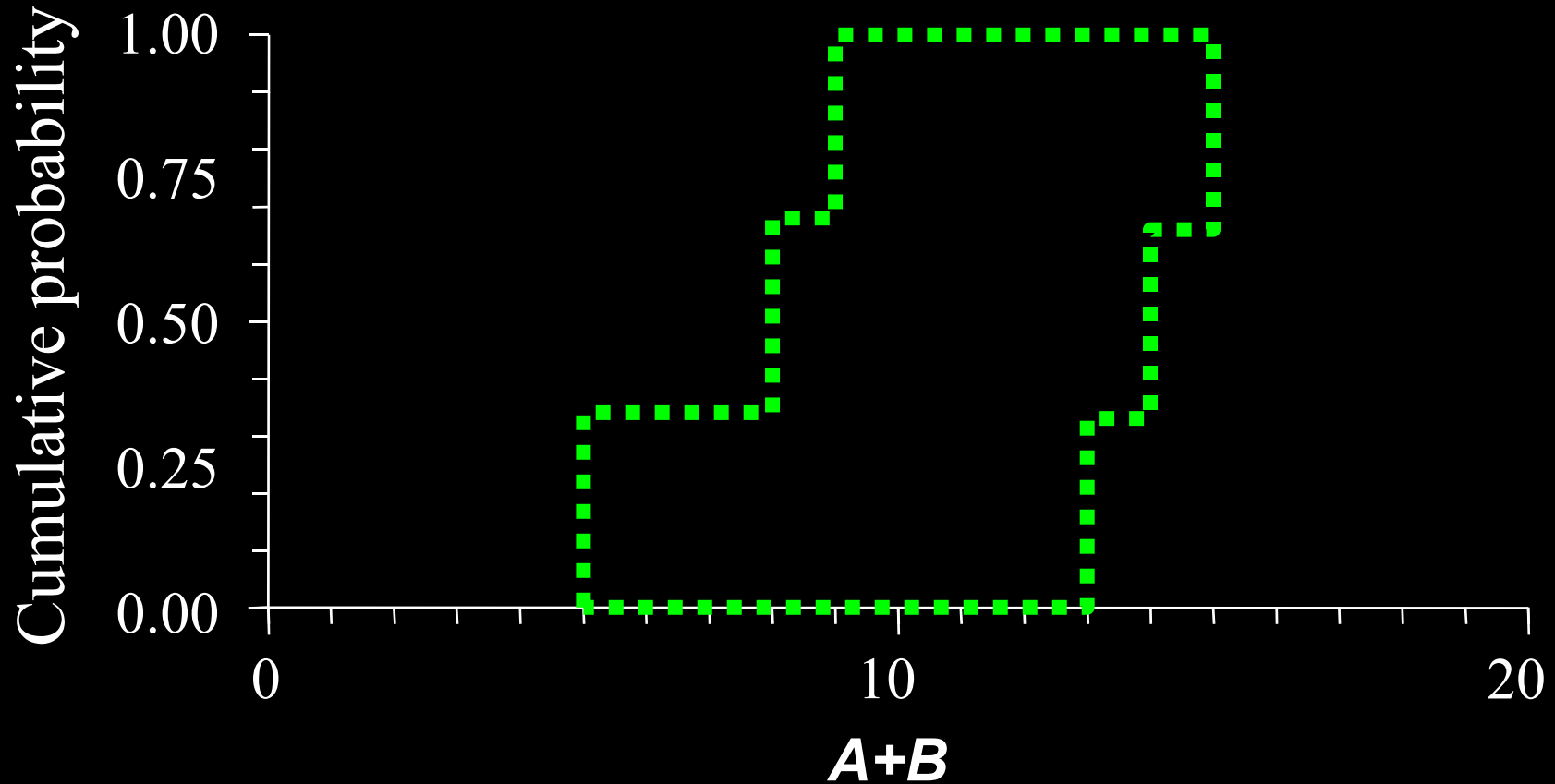
$B \in [8,12]$
 $q_3 = 1/3$

$A+B \in [9,15]$
prob= 1/3

$A+B \in [10,16]$
prob=0

$A+B \in [11,17]$
prob=0

$A+B$, opposite / nondependent



Opposite / opposite

$A+B$

opposite
opposite

$A \in [1,3]$
 $p_1 = 1/3$

$A \in [2,4]$
 $p_2 = 1/3$

$A \in [3,5]$
 $p_3 = 1/3$

$B \in [2,8]$
 $q_1 = 1/3$

$A+B \in [5,9]$
prob=0

$A+B \in [6,10]$
prob=0

$A+B \in [7,11]$
prob=1/3

$B \in [6,10]$
 $q_2 = 1/3$

$A+B \in [9,11]$
prob=0

$A+B \in [10,12]$
prob=1/3

$A+B \in [11,13]$
prob=0

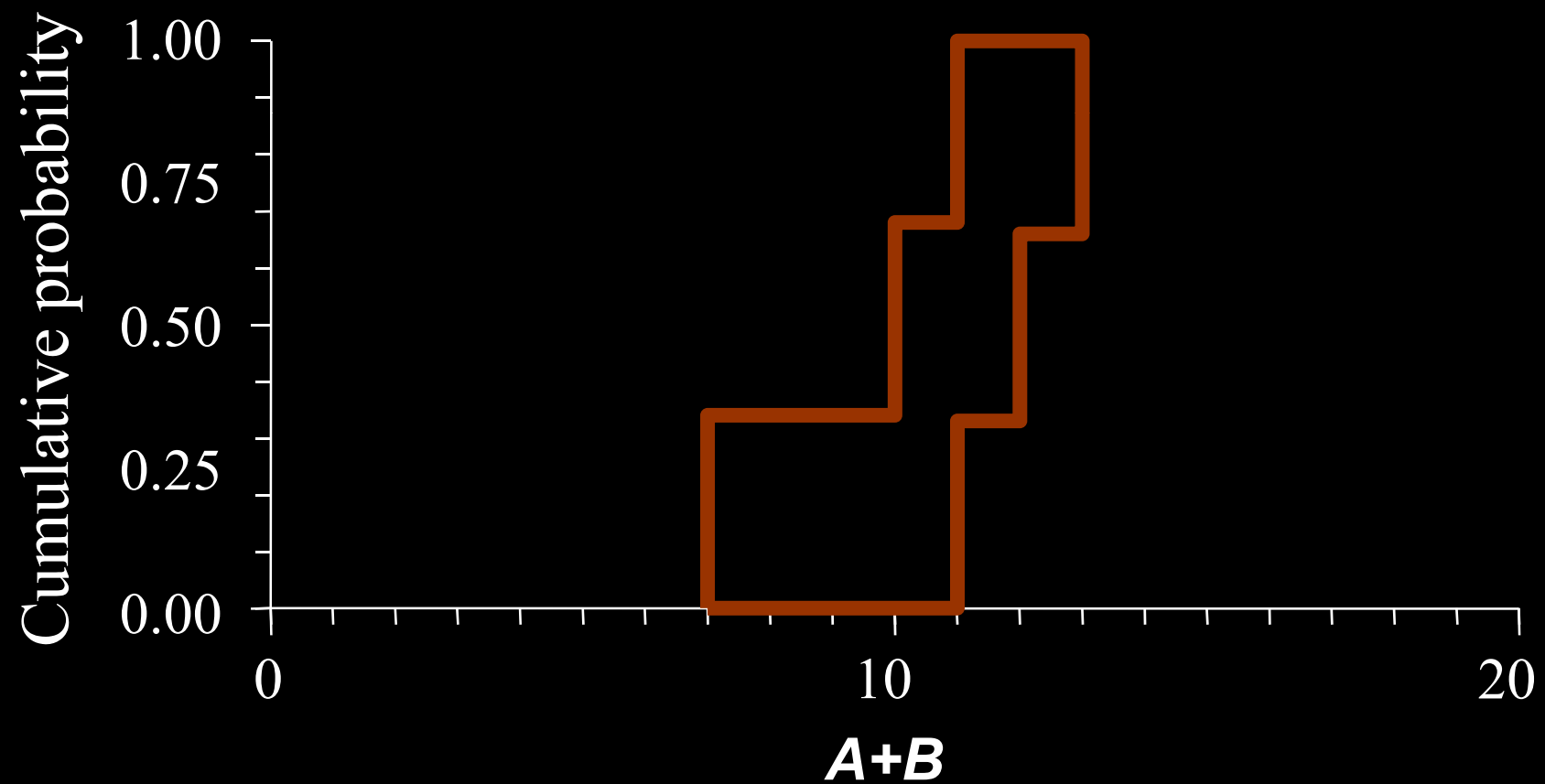
$B \in [8,12]$
 $q_3 = 1/3$

$A+B \in [11,13]$
prob= 1/3

$A+B \in [12,14]$
prob=0

$A+B \in [13,15]$
prob=0

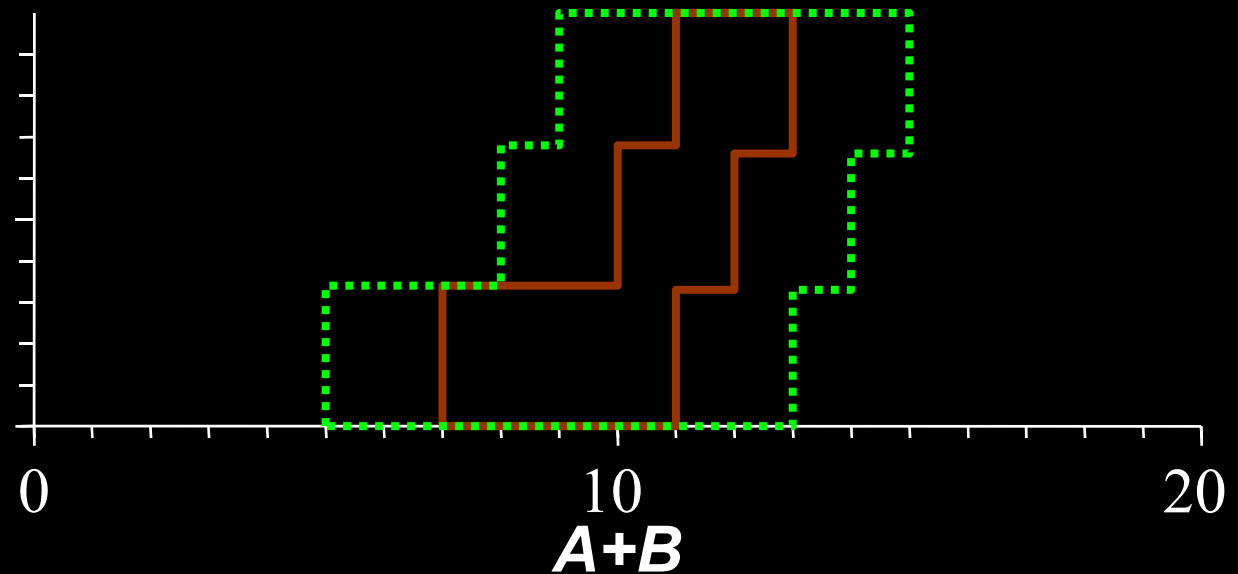
$A+B$, opposite / opposite



Nesting (as we'd expect)

opposite /
nondependent

opposite /
opposite



Two levels and kinds of dependence

- Stochastic dependence
 - *Among* the cells of the Cartesian product
- Interval dependence
 - *Within* each of the cells
- Evidently, they can be considered separately
- Lead to different results in various combinations

Take-home messages

- Interval analysis automatically accounts for *all possible* dependencies
 - Unlike probability theory, where the default assumption often underestimates uncertainty
- Information about dependencies isn't usually used to tighten results, *but it can be*
- Interval and stochastic dependence are distinct, and may interact to tighten results

When p-boxes won't do

What bounding probability can't do

- Represent **comparative probability judgments**, e.g., event A is at least as likely as event B
- Give unique **expectations** needed for making decisions
- Give unique **conditional probabilities** needed for making inferences
- Maintain best possible bounds through **updating**

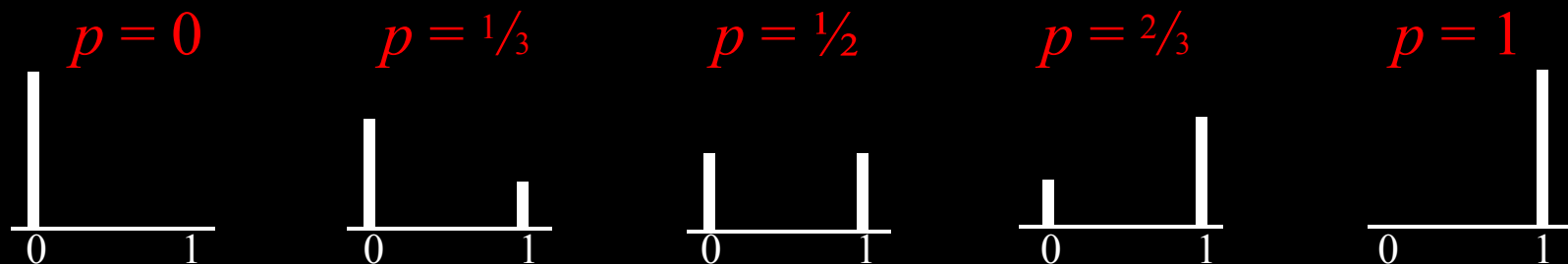
(Walley 2000)

Imprecise probabilities

- This is really a generic term referring to the subject matter of many disparate theories that do not assume a unique underlying probability measure
- Often expressed in a lower probability measure which gives the lower probability for every possible event in some universal set
- Often expressed in terms of closed, convex sets of probability distributions (not the same as a p-box)

Sets of distribution functions

- Consider the set of all Bernoulli distributions (which are discrete with mass at only 0 and 1)



- Clearly, there's a one-dimensional family of such distributions, parameterized by how the mass is distributed between the two points

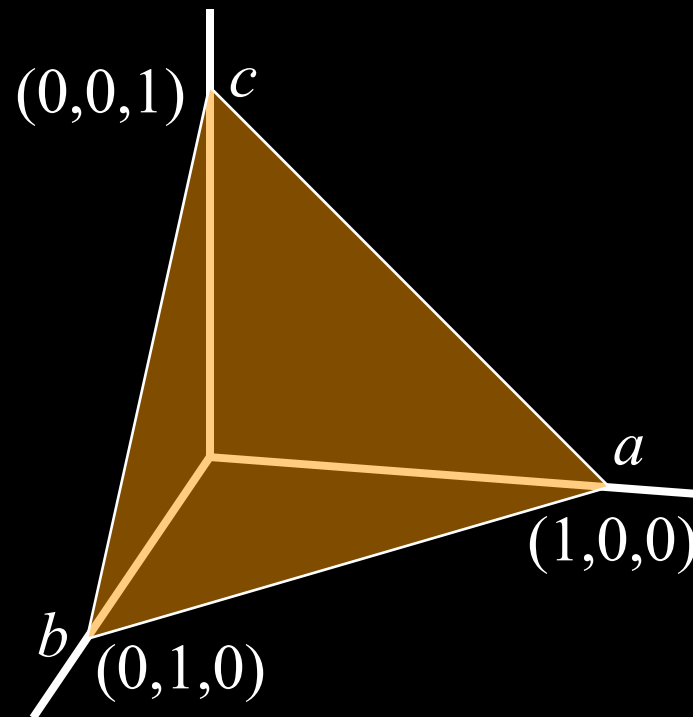
Space of distributions

- This one-dimensional family constitutes a space of distributions in which each point represents a distribution



Three-dimensional case

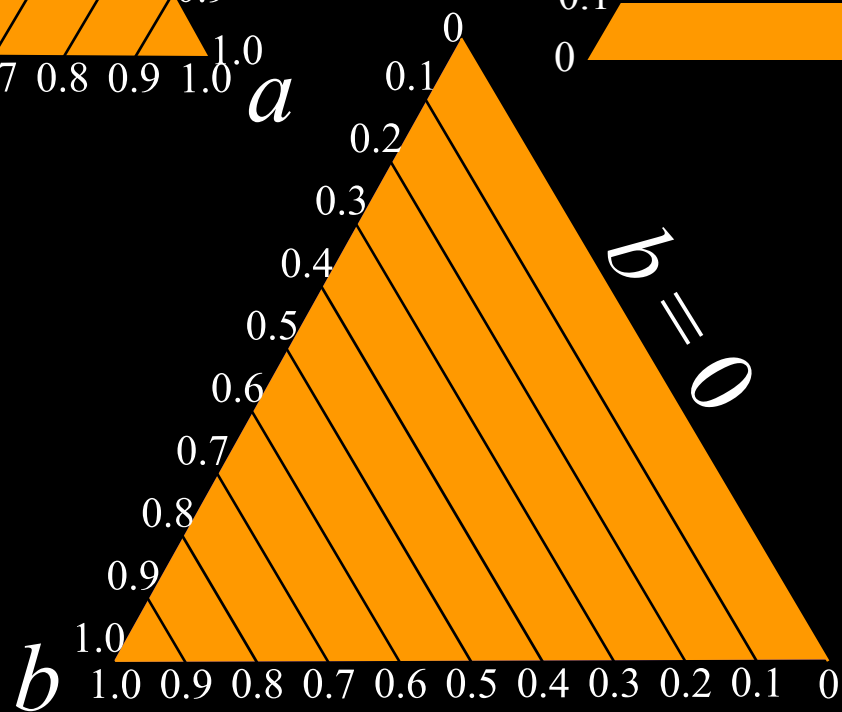
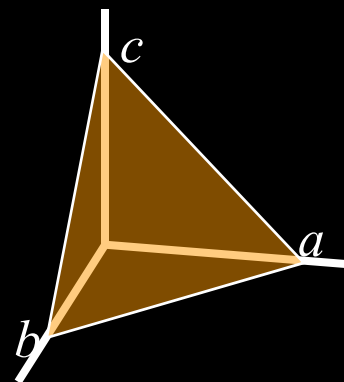
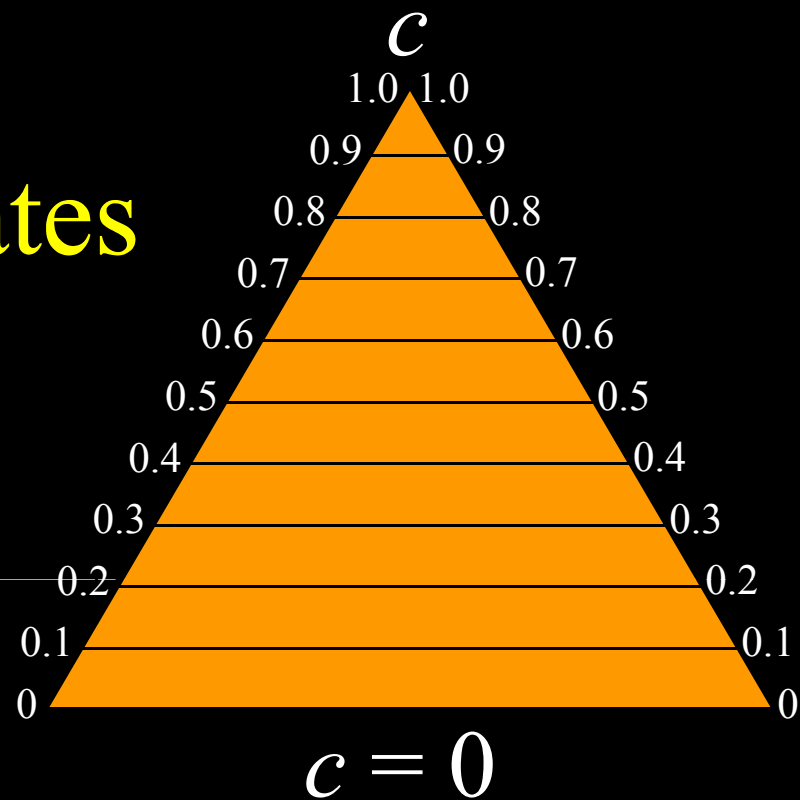
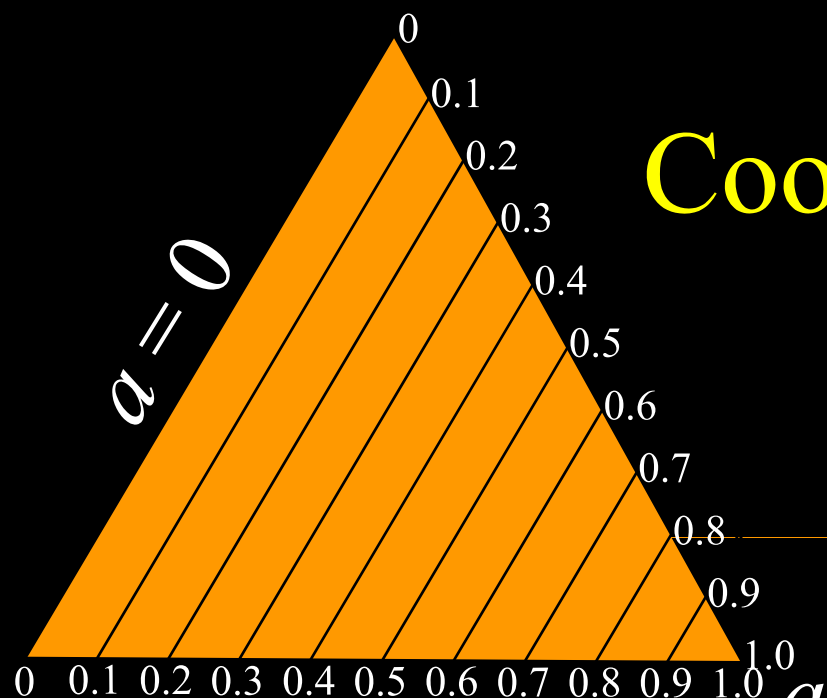
- When the distributions have 3 point masses, the space becomes two-dimensional and has a triangular shape
- The points on this surface are those whose coordinates add to one



Simplex

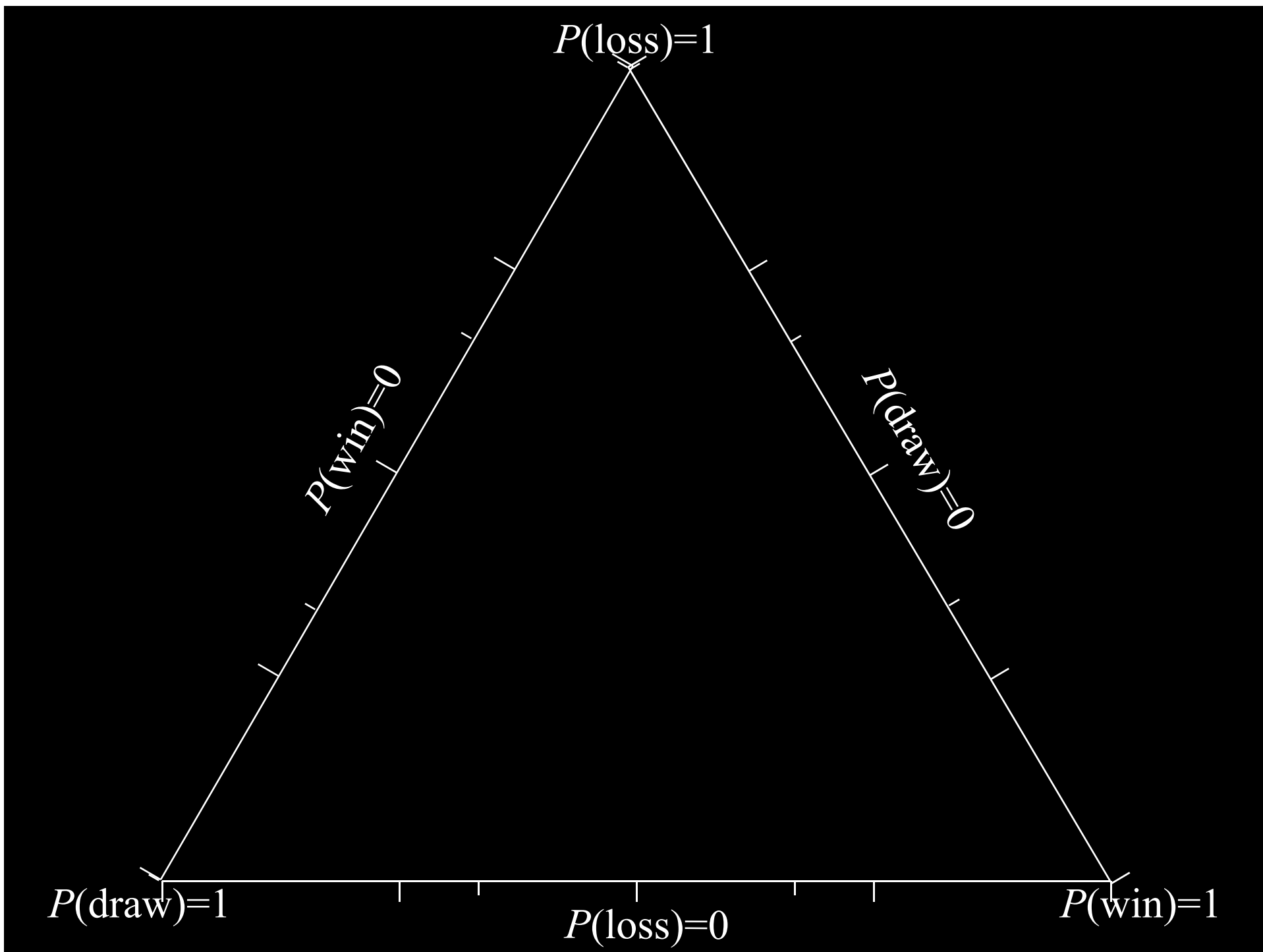
- For discrete distributions with n masses, the space, called a *simplex*, has $(n-1)$ -dimensions
- One degree of freedom is lost to the constraint that probabilities sum to one
- For the continuous case, the space becomes infinite-dimensional, or you could be content to use discrete approximations

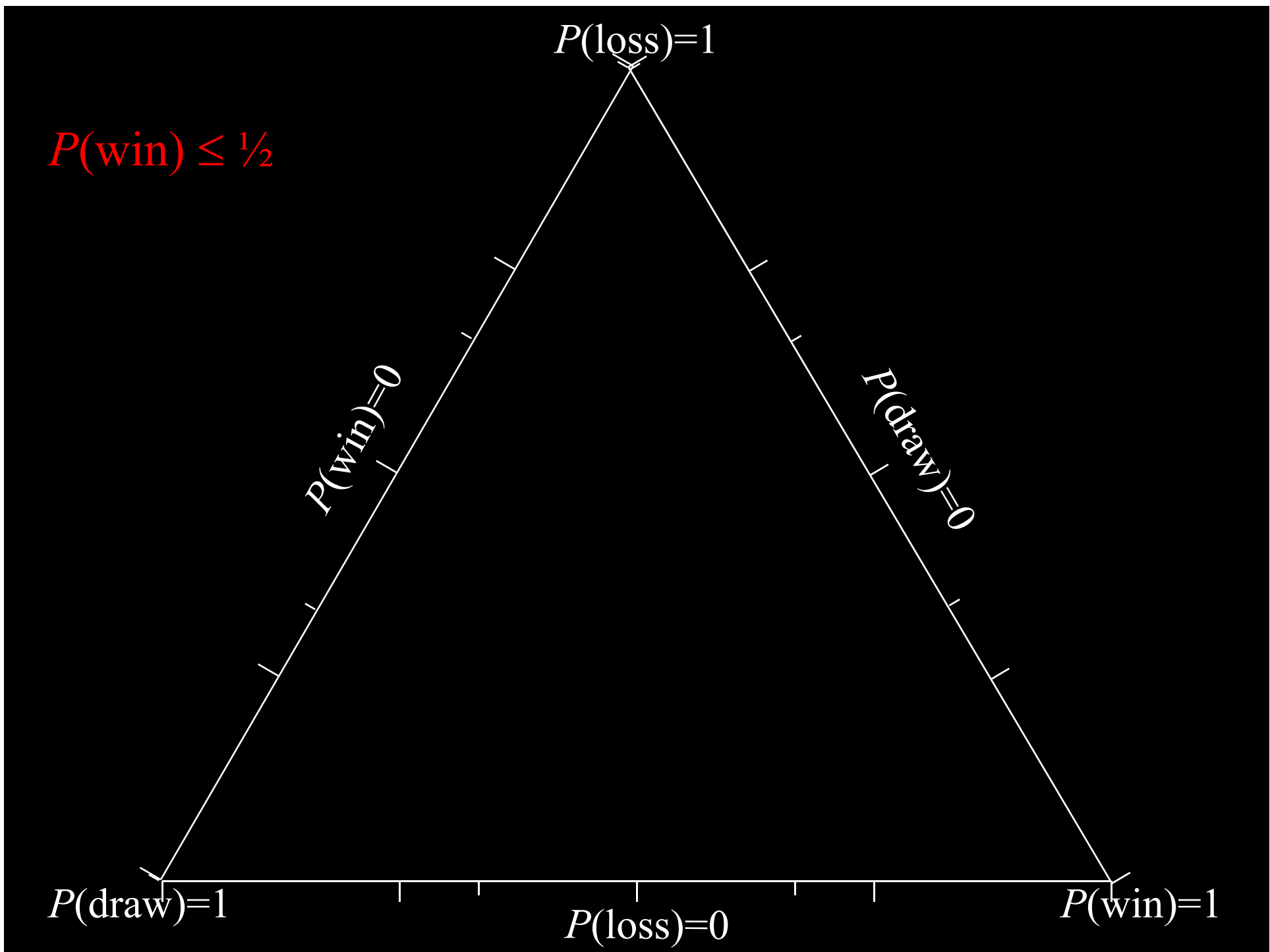
Coordinates

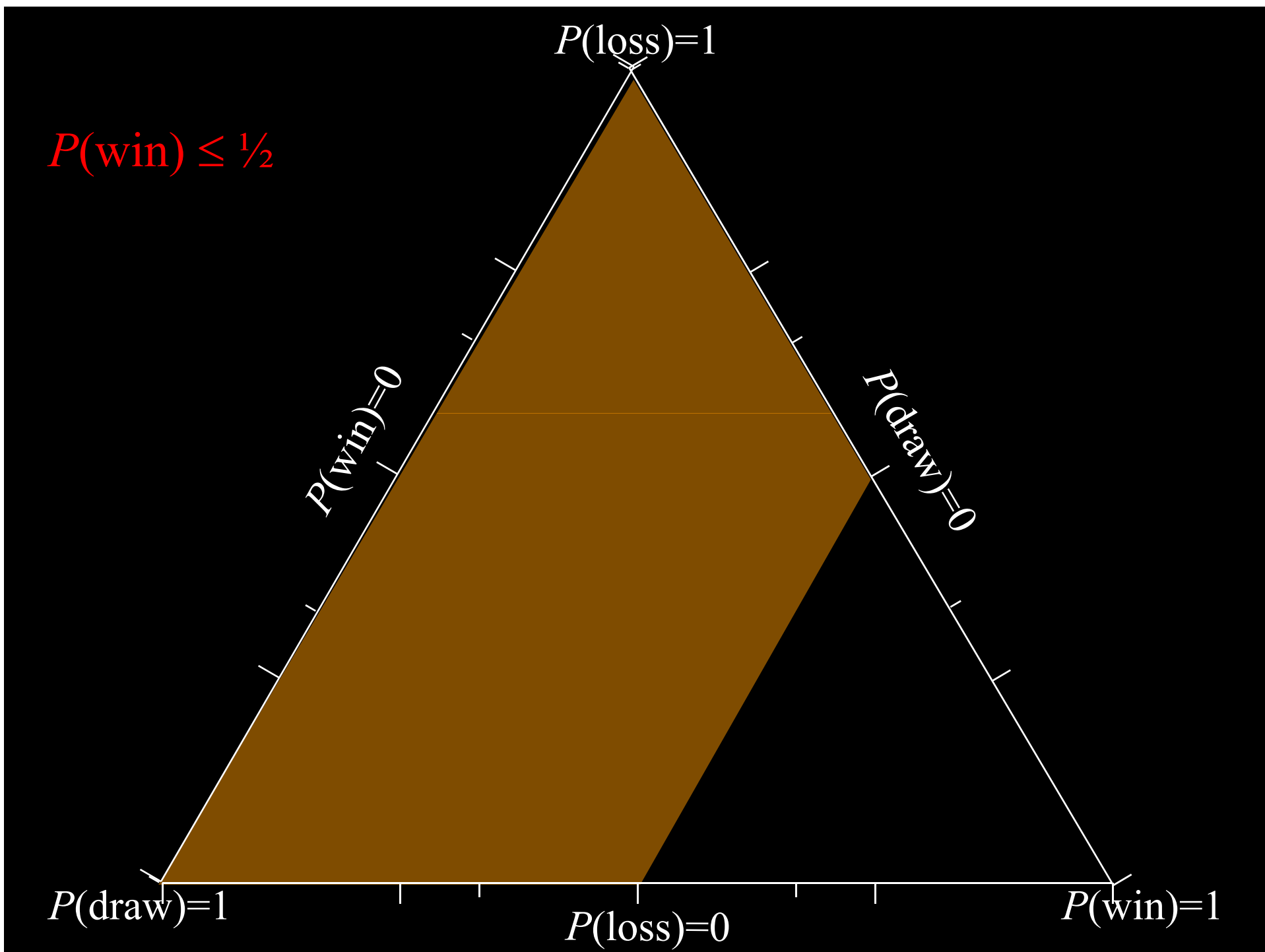


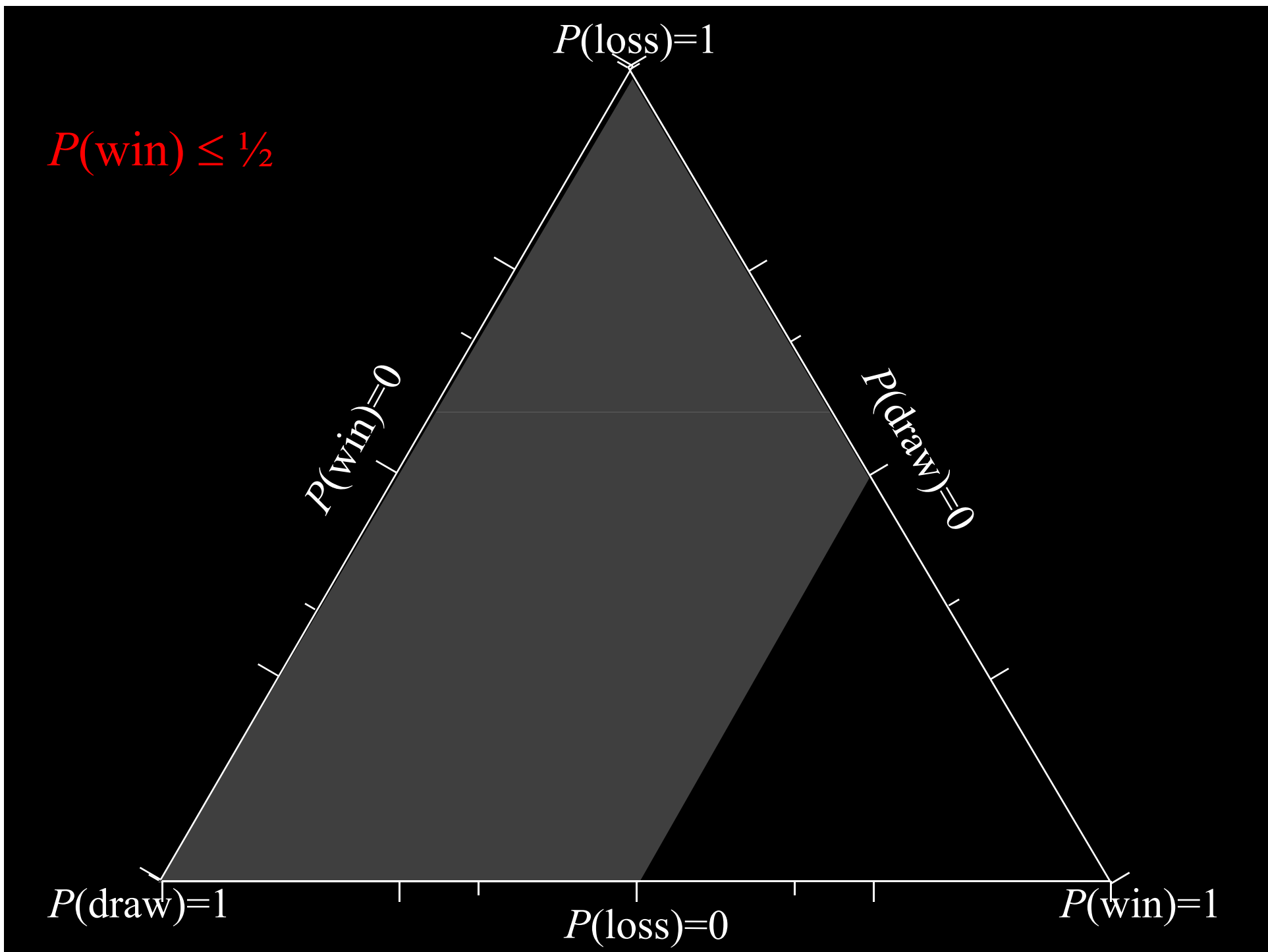
Walley's (2000) football game

- 3 possibilities for our team: win, draw, loss
- Suppose we have qualitative judgments:
 - ‘Not win’ is at least as probable as win
 - Win is at least as probable as draw
 - Draw is at least as probable as loss
- These constrain the probability distribution P
 - $P(\text{win}) \leq \frac{1}{2}$
 - $P(\text{win}) \geq P(\text{draw})$
 - $P(\text{draw}) \geq P(\text{loss})$









$P(\text{loss})=1$

$P(\text{win}) \leq \frac{1}{2}$

$P(\text{draw}) \geq P(\text{loss})$

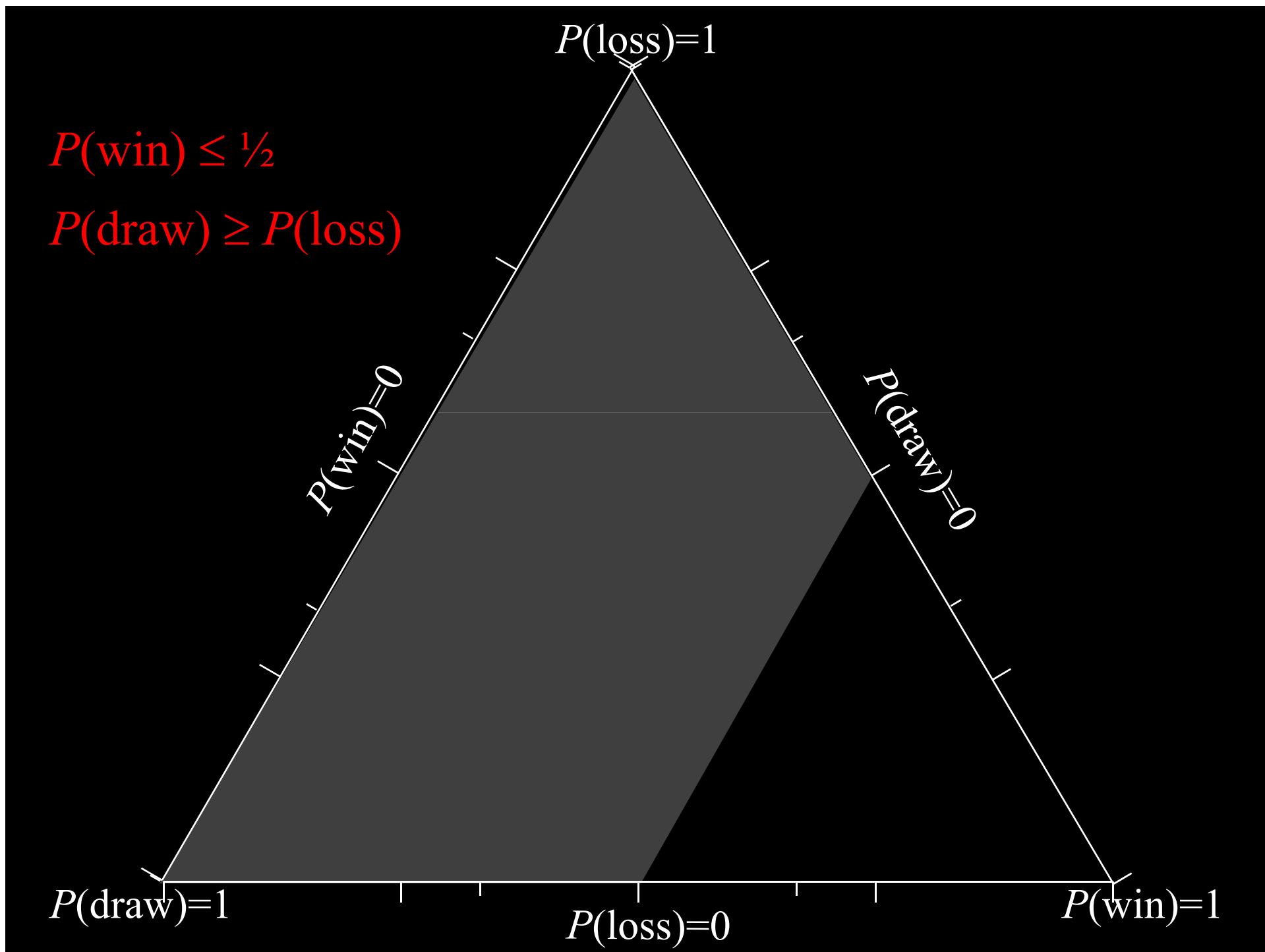
$P(\text{win})=0$

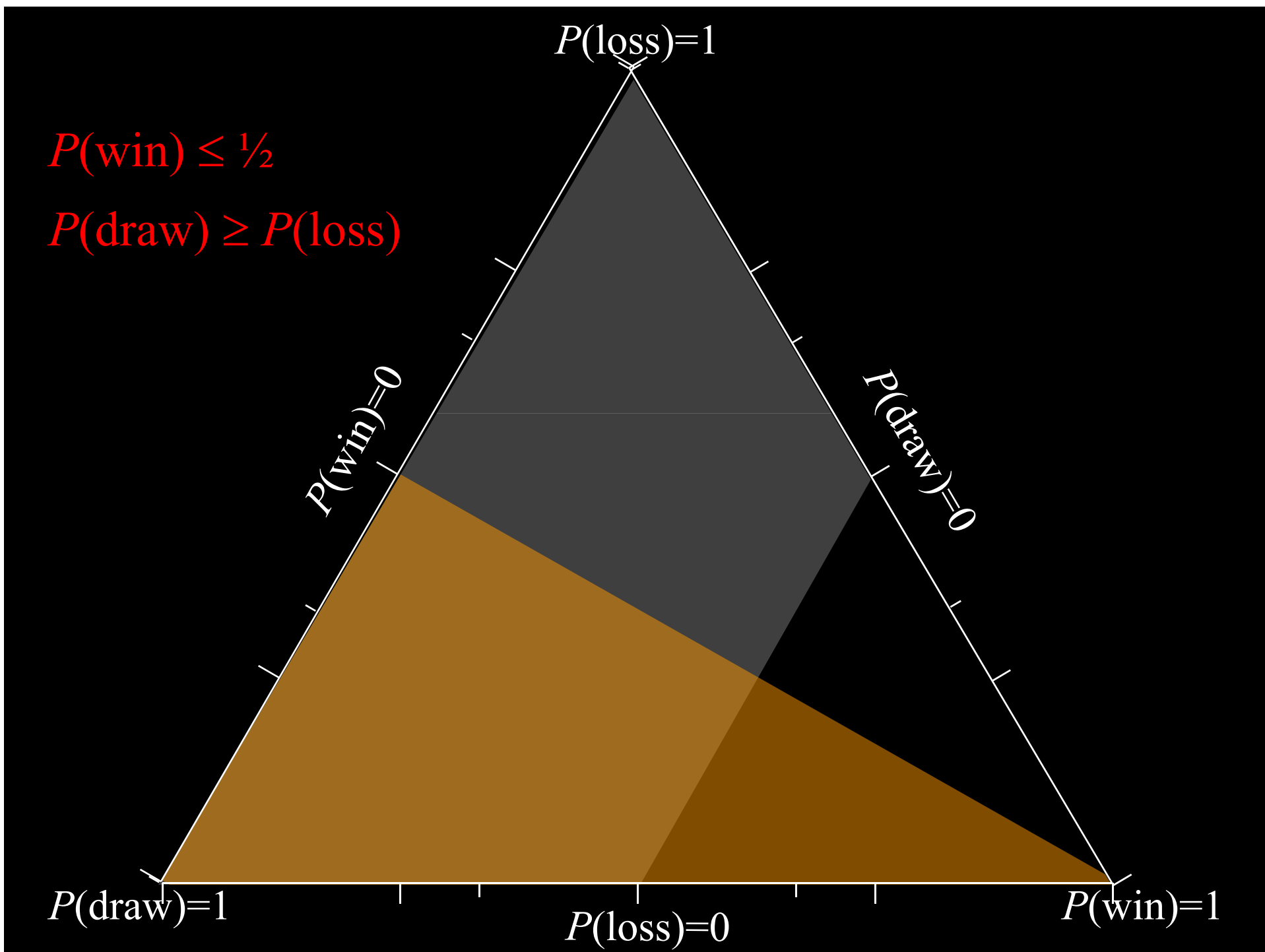
$P(\text{draw})=0$

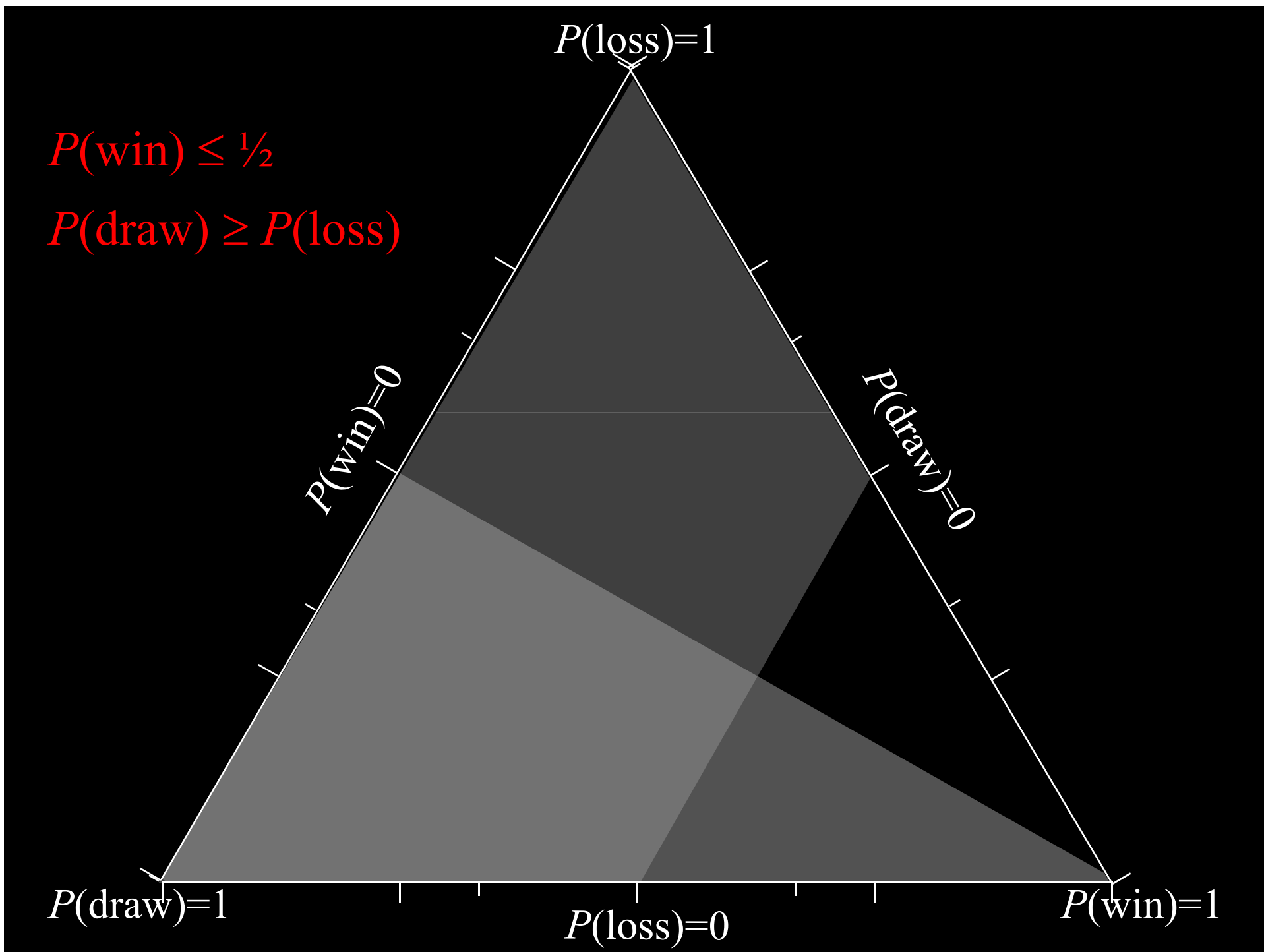
$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$







$P(\text{loss})=1$

$P(\text{win}) \leq \frac{1}{2}$

$P(\text{draw}) \geq P(\text{loss})$

$P(\text{win}) \geq P(\text{draw})$

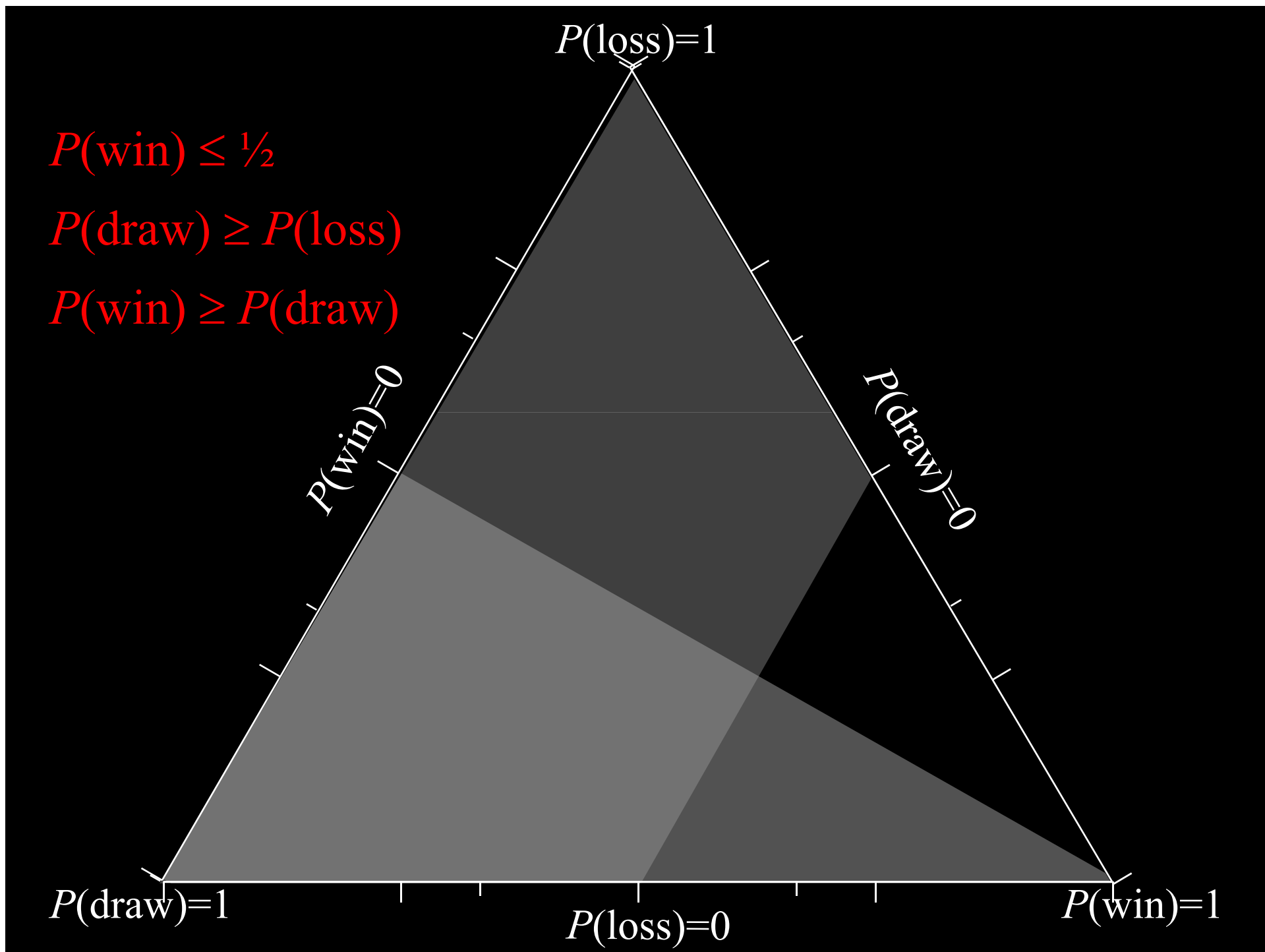
$P(\text{win})=0$

$P(\text{draw})=0$

$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$



$P(\text{loss})=1$

$$P(\text{win}) \leq \frac{1}{2}$$

$$P(\text{draw}) \geq P(\text{loss})$$

$$P(\text{win}) \geq P(\text{draw})$$

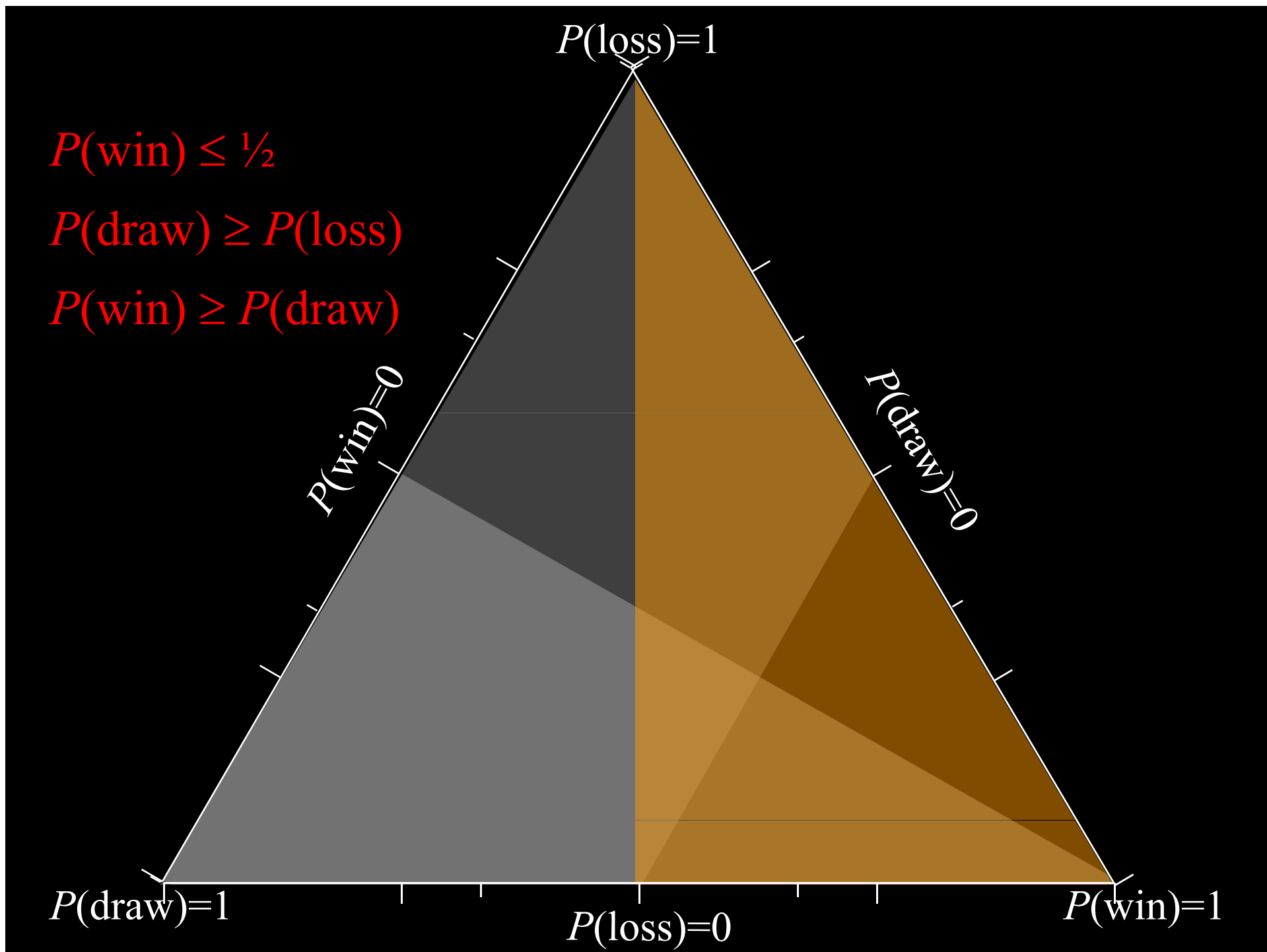
$P(\text{win})=0$

$P(\text{draw})=0$

$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$



$P(\text{loss})=1$

$P(\text{win}) \leq \frac{1}{2}$

$P(\text{draw}) \geq P(\text{loss})$

$P(\text{win}) \geq P(\text{draw})$

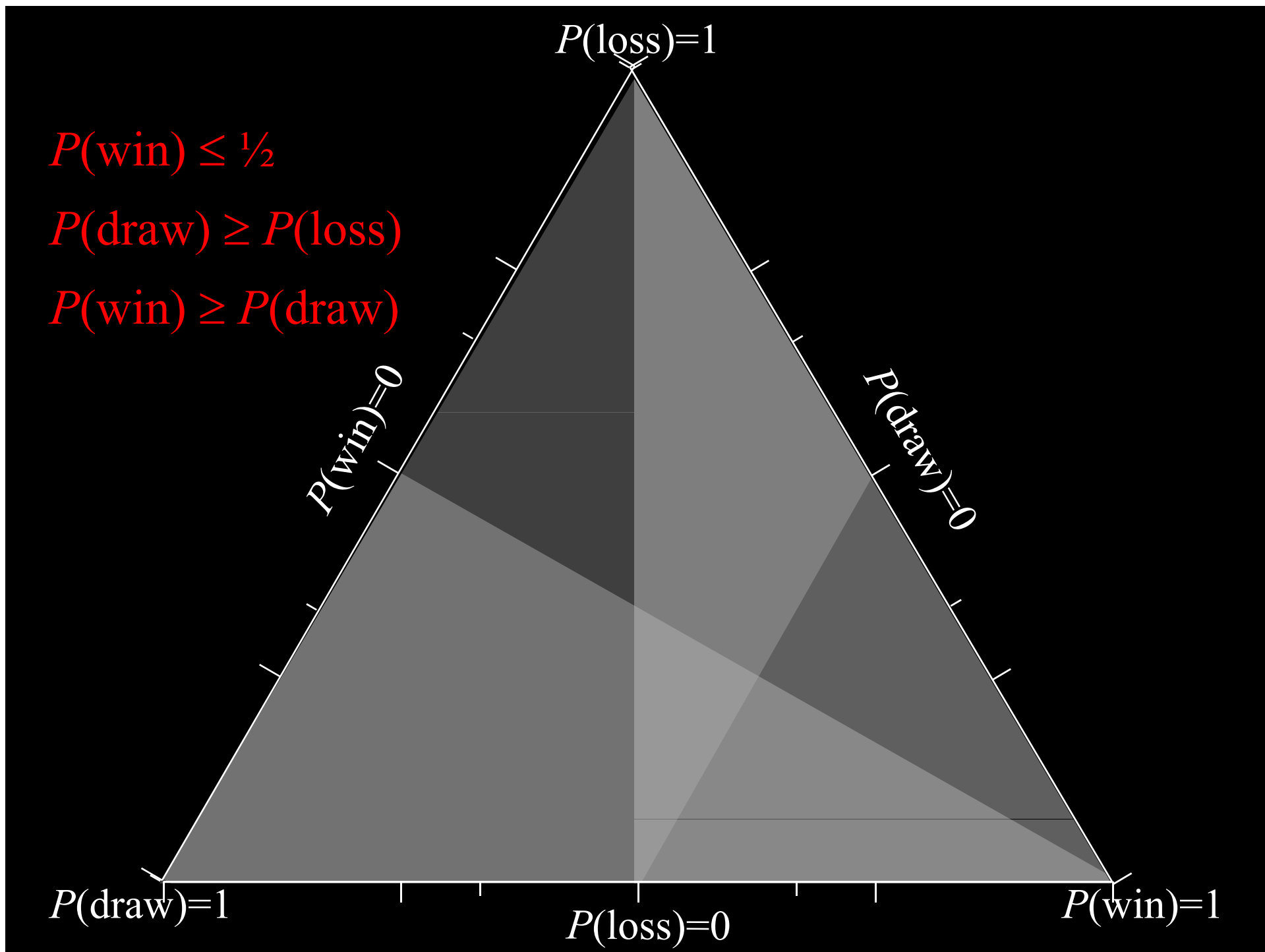
$P(\text{win})=0$

$P(\text{draw})=0$

$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$



$P(\text{loss})=1$

$P(\text{win}) \leq \frac{1}{2}$

$P(\text{draw}) \geq P(\text{loss})$

$P(\text{win}) \geq P(\text{draw})$

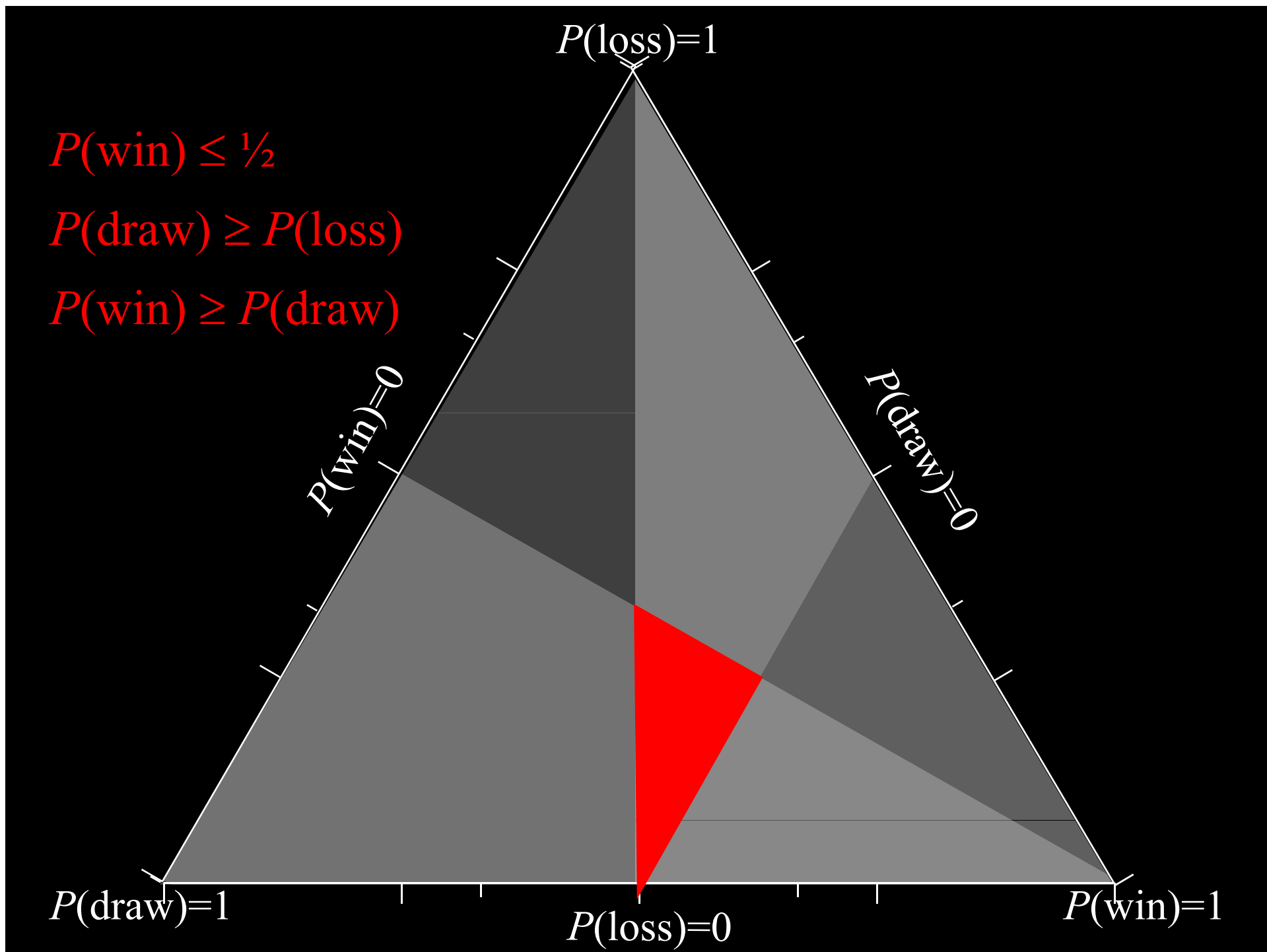
$P(\text{win})=0$

$P(\text{draw})=0$

$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$



$P(\text{loss})=1$

$P(\text{win}) \leq \frac{1}{2}$

$P(\text{draw}) \geq P(\text{loss})$

$P(\text{win}) \geq P(\text{draw})$

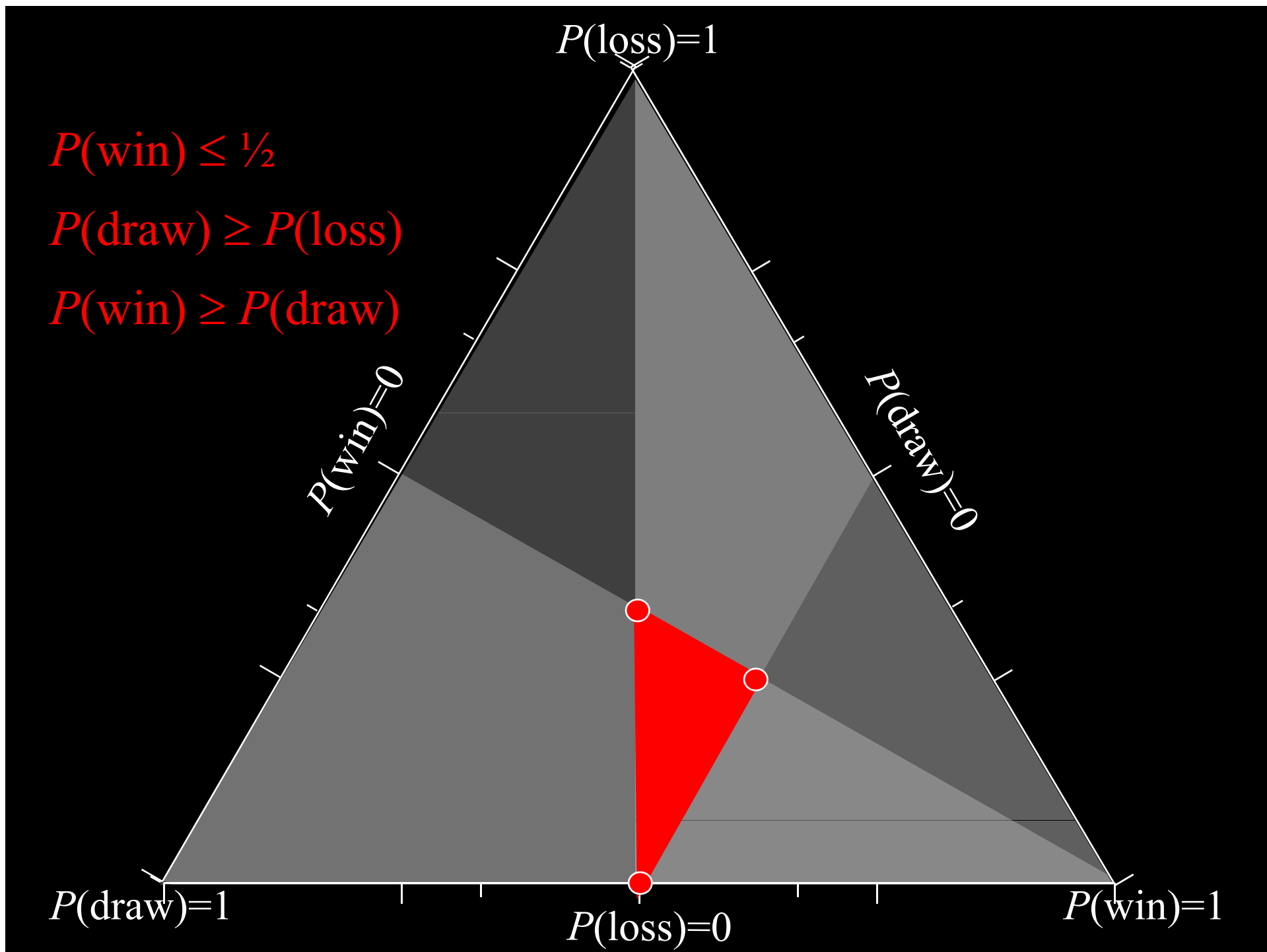
$P(\text{win})=0$

$P(\text{draw})=0$

$P(\text{draw})=1$

$P(\text{loss})=0$

$P(\text{win})=1$



Natural extension

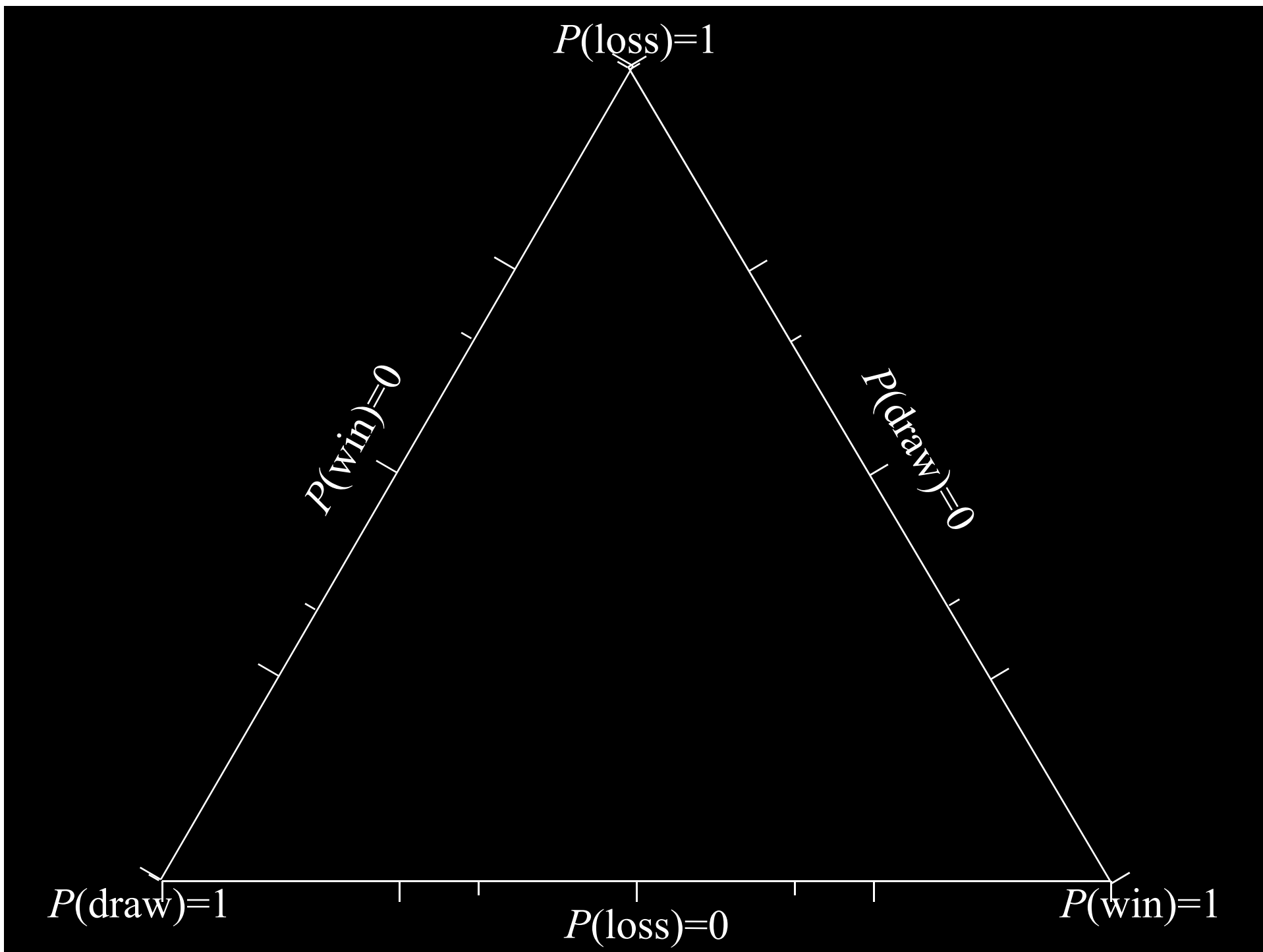
- Natural extension tells us the extreme points (obtained by linear programming) that are consistent with these constraints are

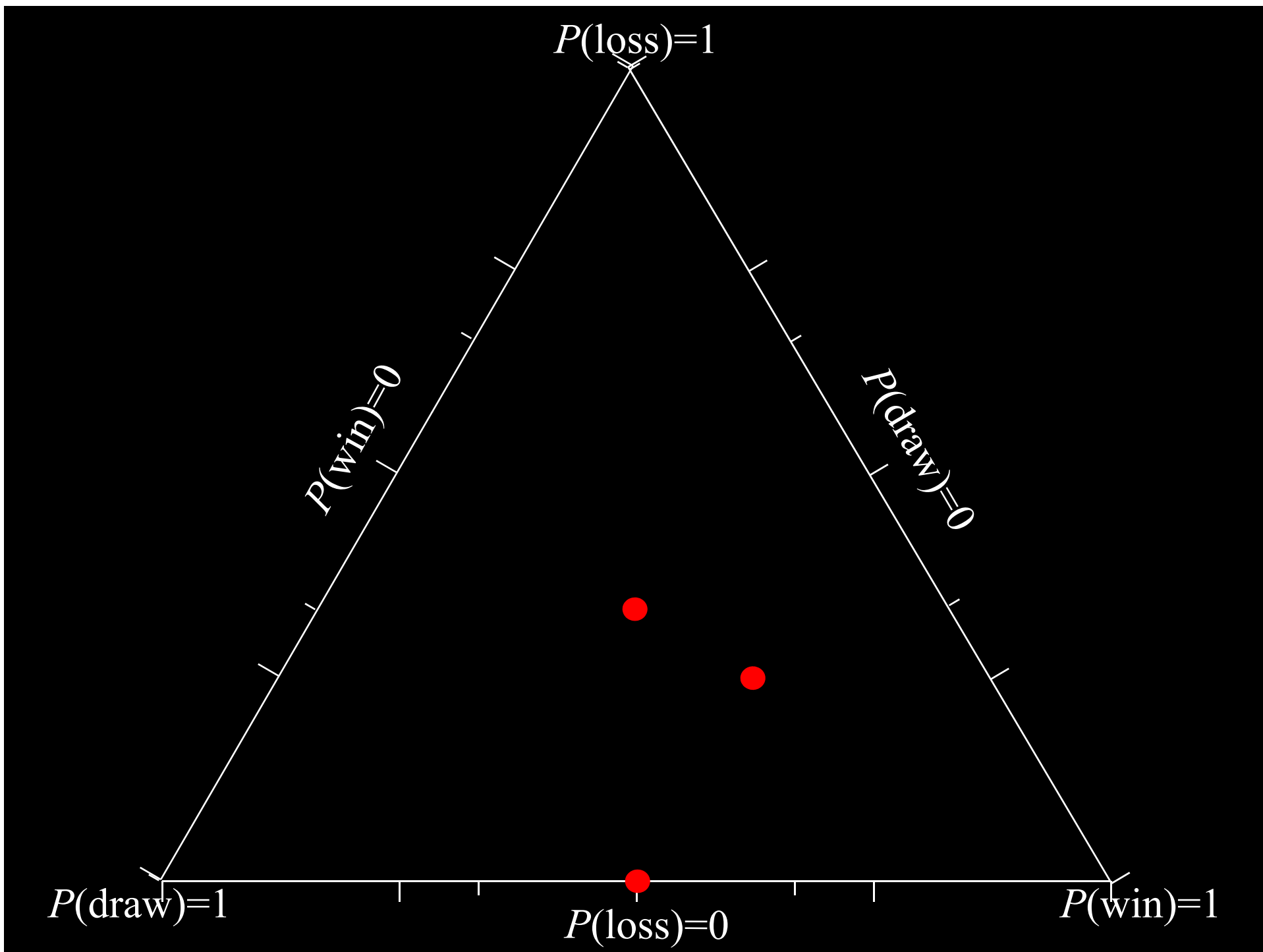
$$(\frac{1}{2}, \frac{1}{2}, 0)$$

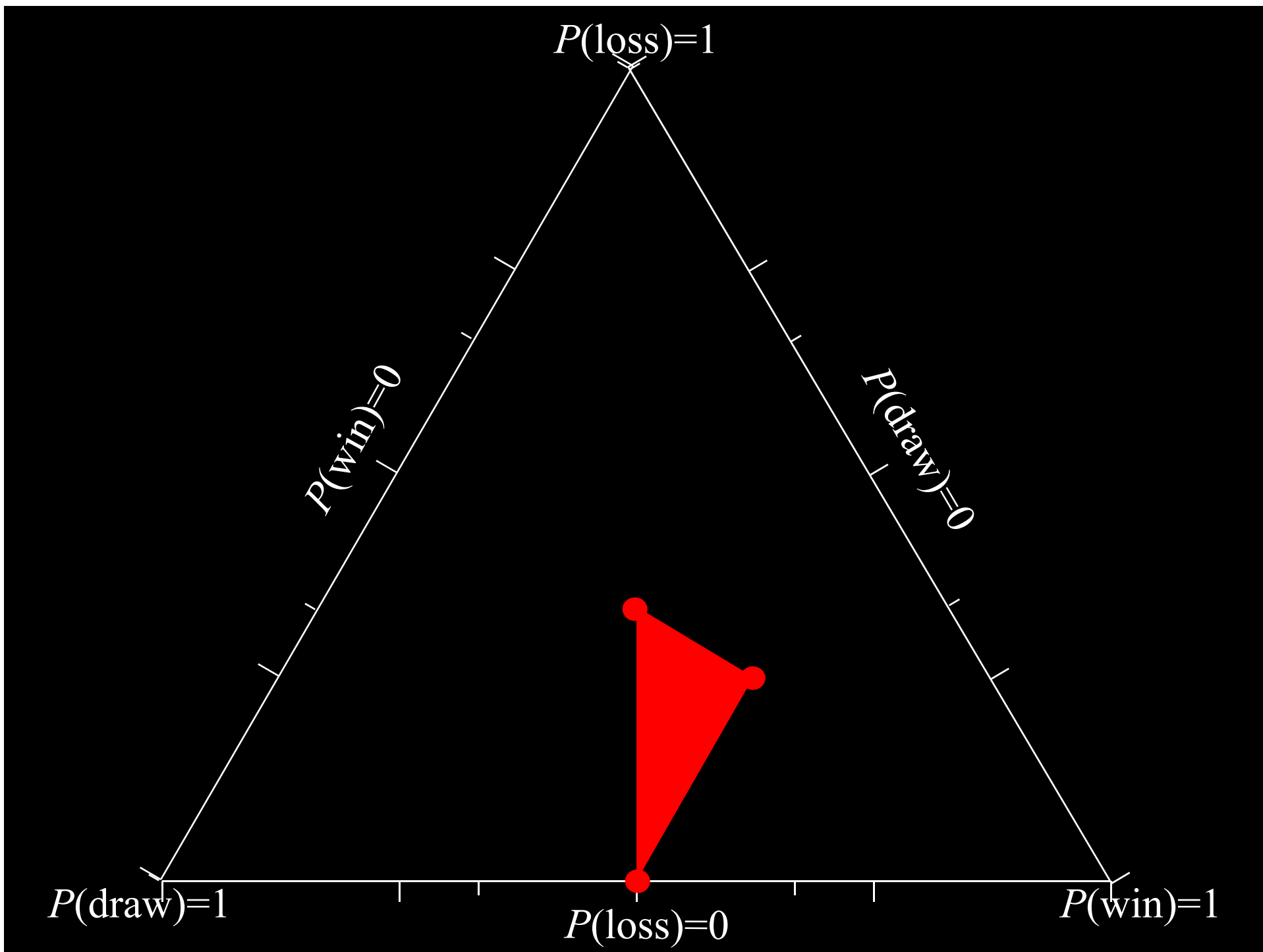
$$(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$$

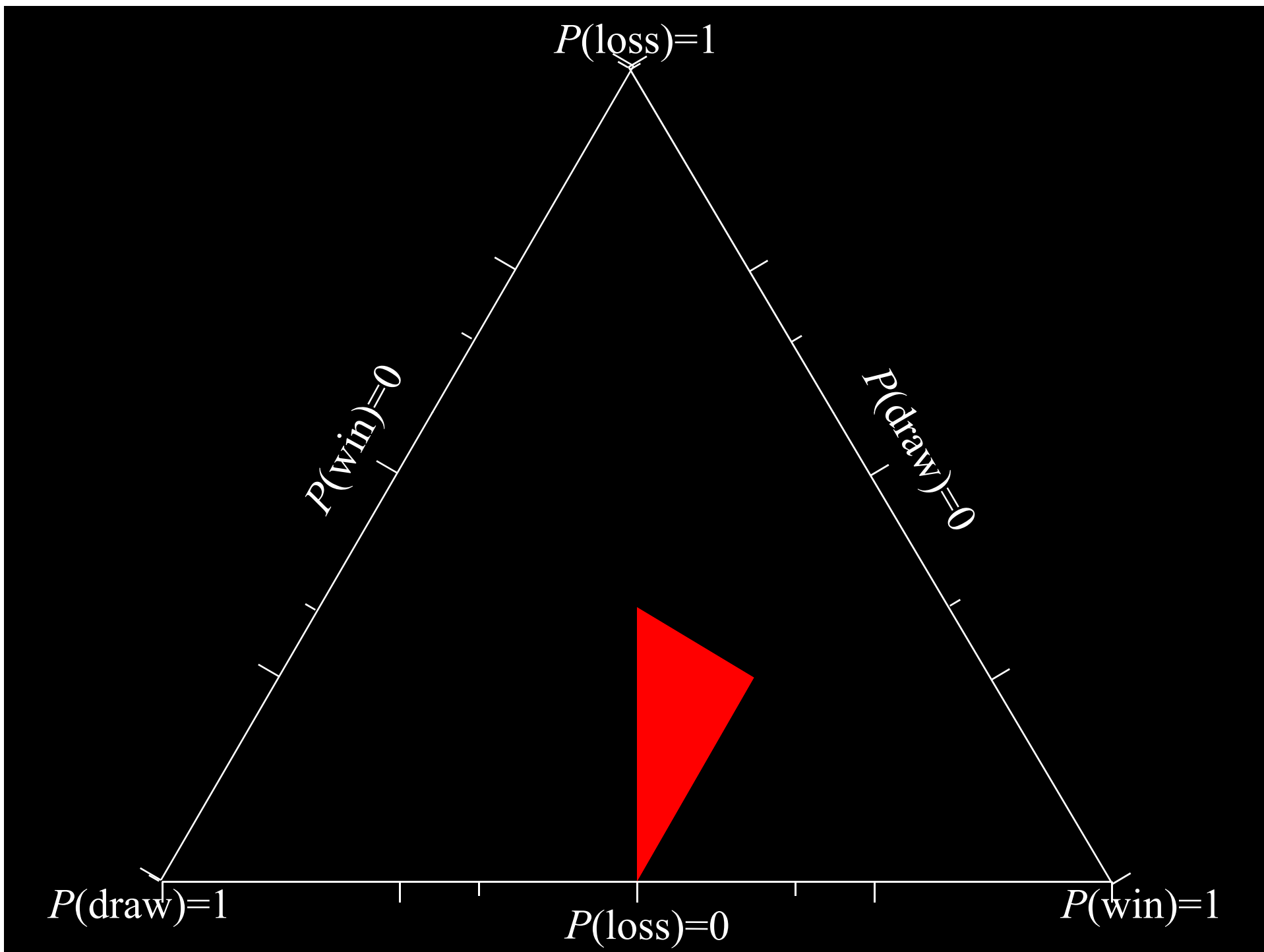
$$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

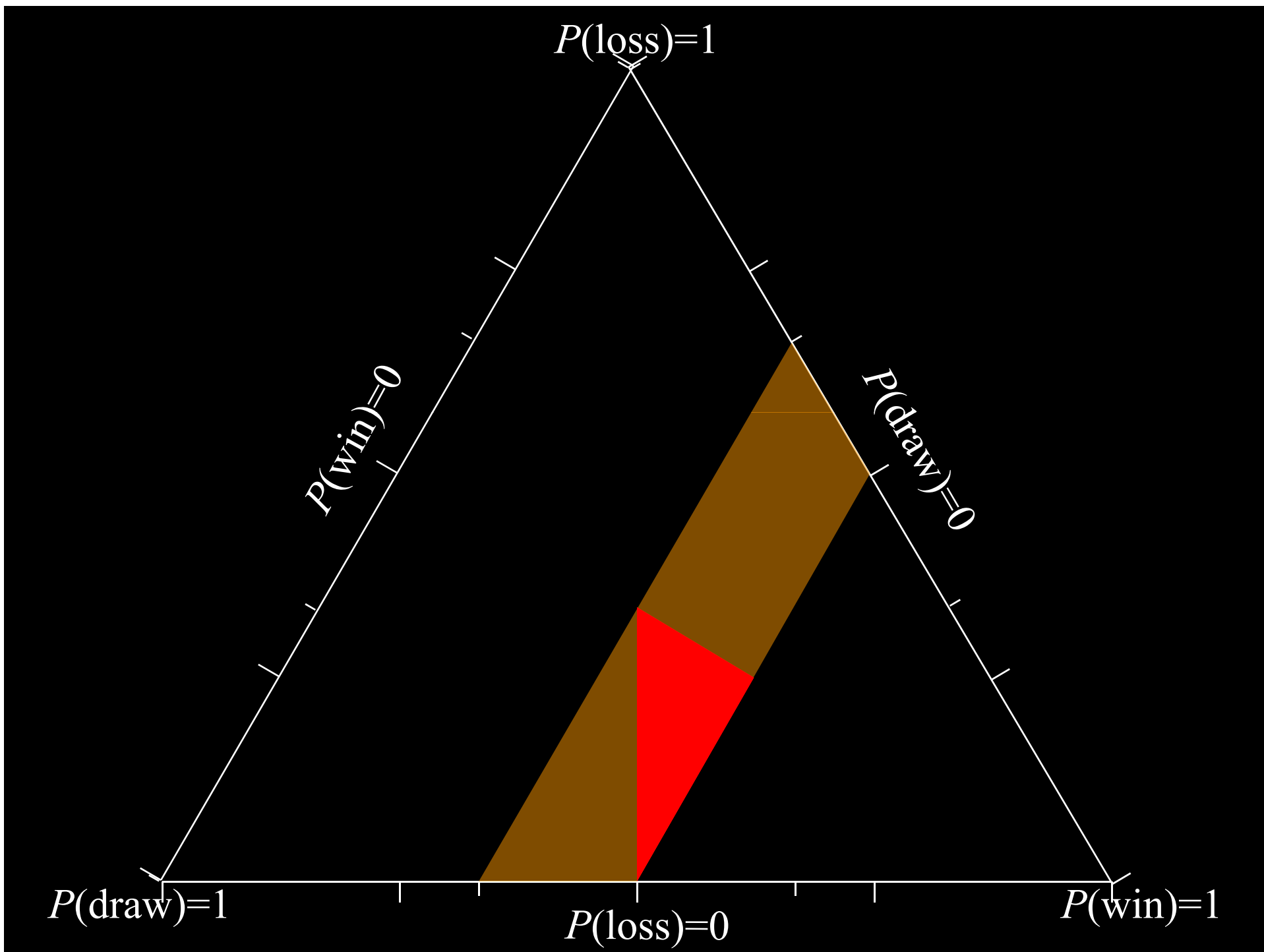
- The points represent extreme distributions
- Their convex hull gives all distributions that are consistent with the constraints

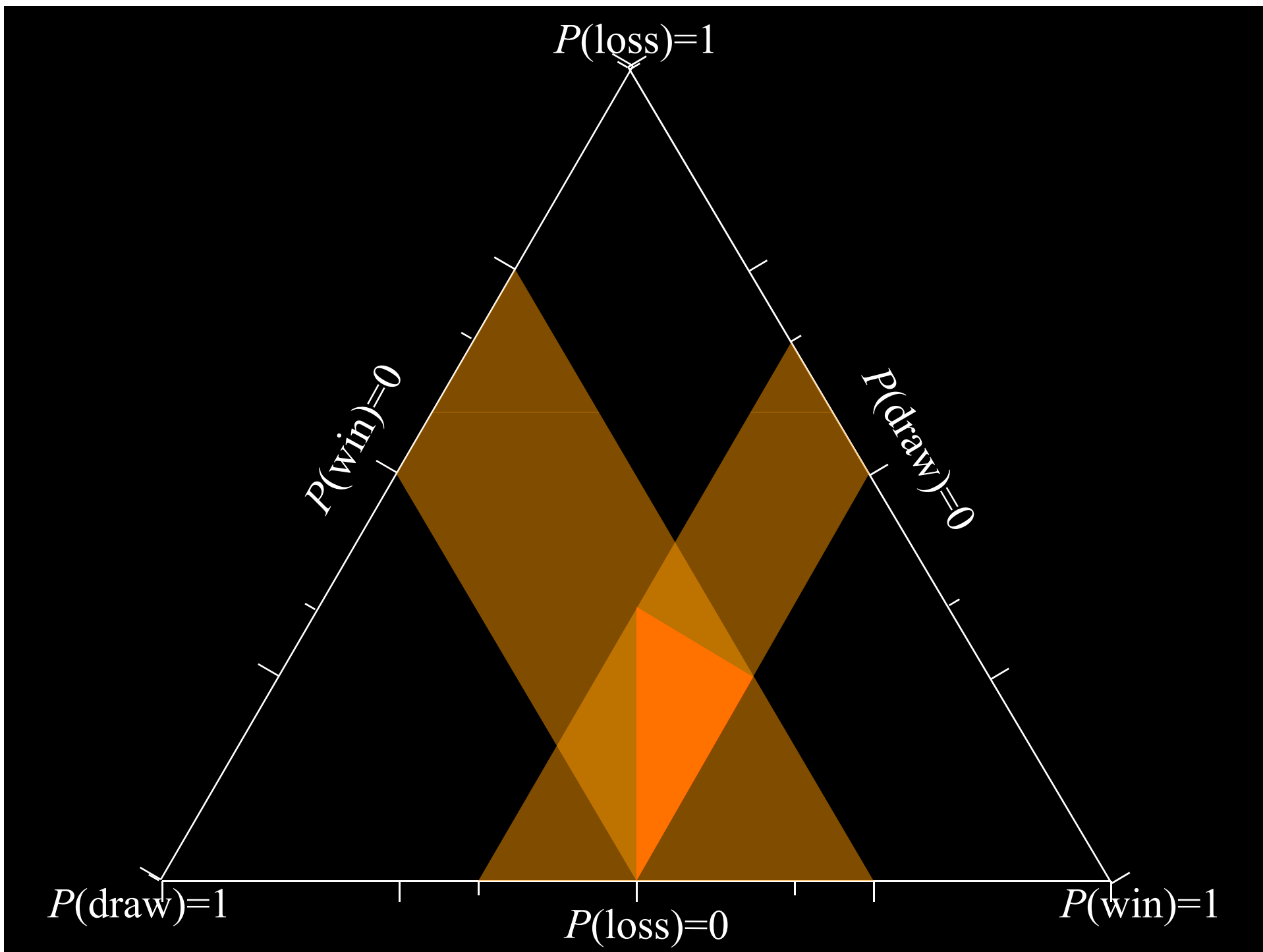


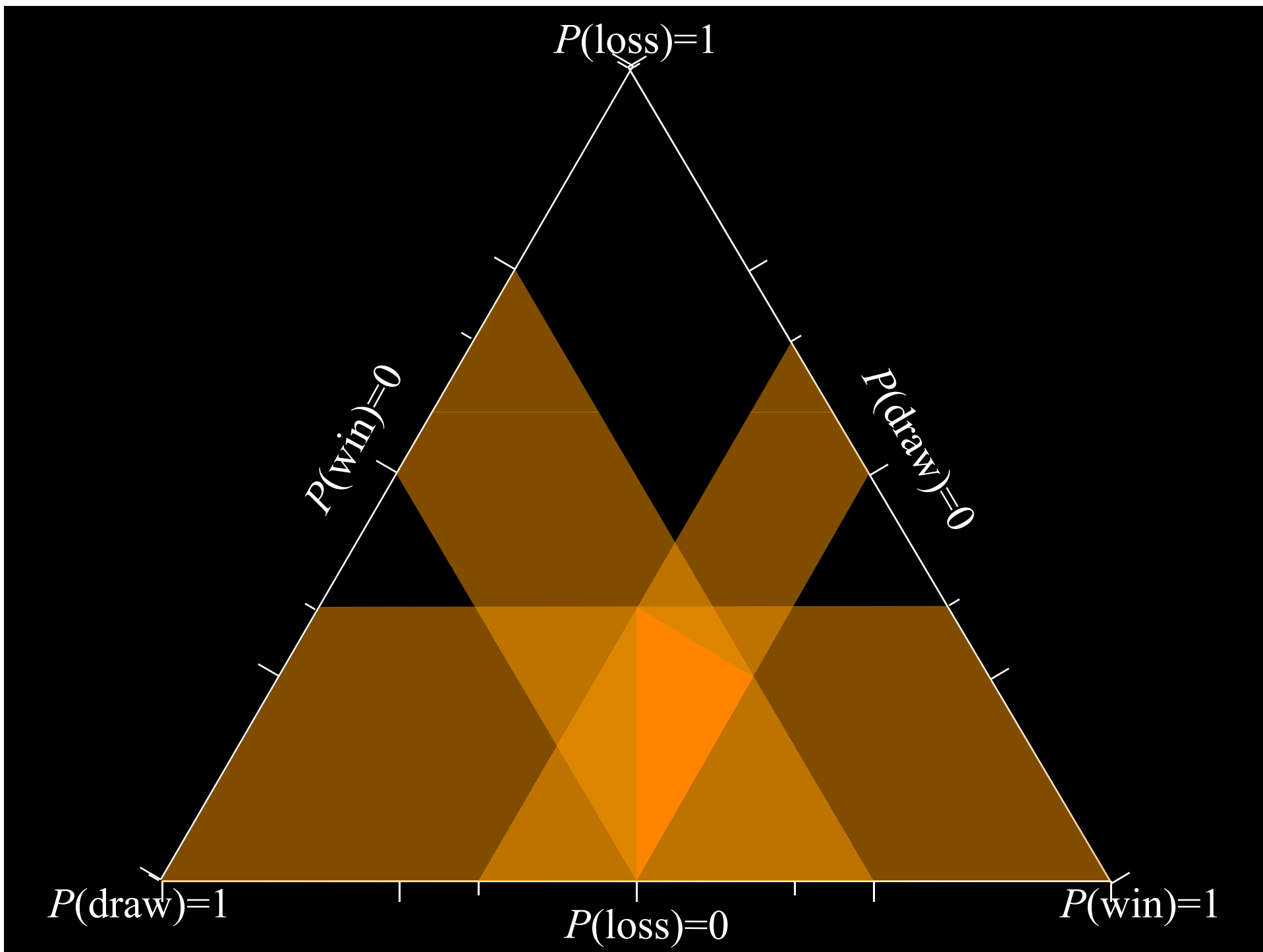


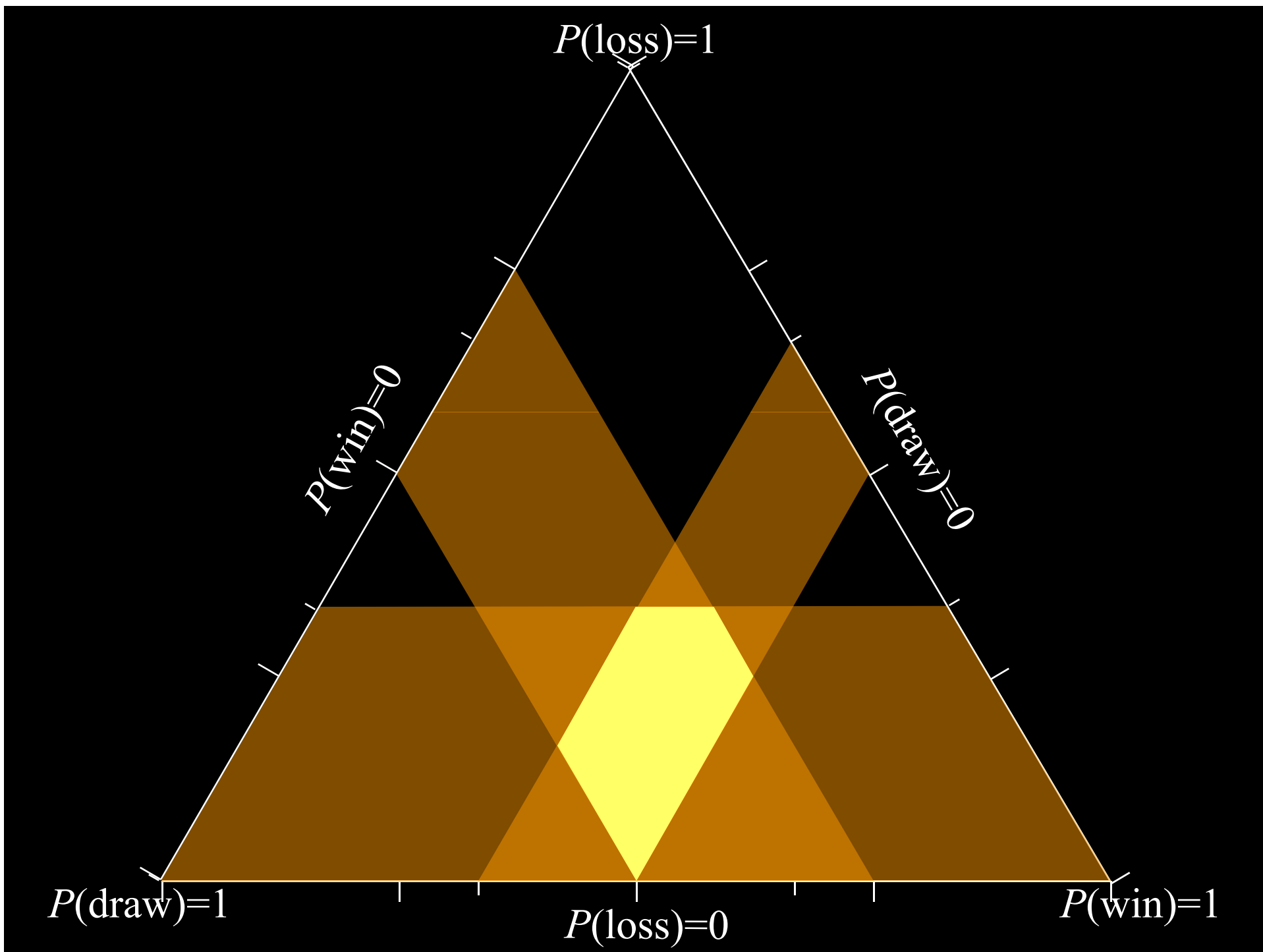


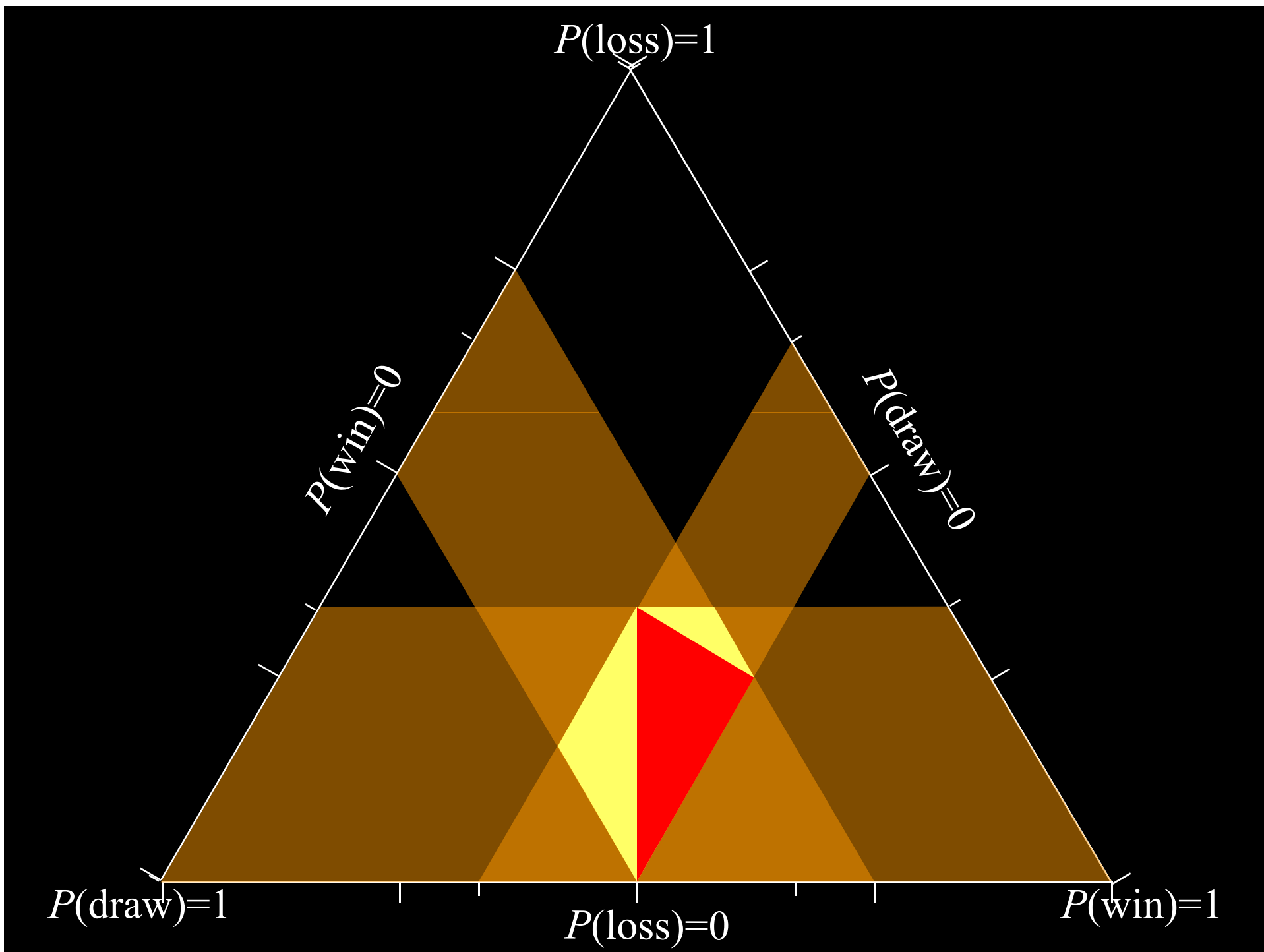












So what?

- The closed, convex set of probability distributions (the red triangular region) expresses the uncertainty
- This set of distributions is **smaller** than the set implied by bounds on the three probabilities (the yellow area enclosing the triangle)
- This difference can affect expectations of functions that depend on the events, and conditional probabilities

Credal set

- Knowledge and judgments can be used to define a set of possible probability measures
 - All distributions within bounds are possible
 - Only distributions having a given shape
 - Probability of an event is within some interval
 - Event A is at least as probable as event B
 - Nothing is known about the probability of C

Take-home lesson

- As was true for interval probabilities and probability bounds, it can be easy to get **rigorous** bounds, but **best possible** bounds may require mathematical programming

Independence in IP

Independence

- In the context of precise probabilities, there was a unique notion of independence
- In the context of imprecise probabilities, however, this notion disintegrates into several distinct concepts (Couso et al. 1999)
- The different kinds of independence behave differently in computations (Fetz 2001)

Equivalent definitions of independence

For precise probabilities, all these definitions are equivalent, so there's a *single concept*

- $H(x,y) = F(x) G(y)$, for all values x and y
- $P(X \in I, Y \in J) = P(X \in I) P(Y \in J)$, for any $I, J \subseteq \mathbf{R}$
- $h(x,y) = f(x)g(y)$, for all values x and y
- $E(w(X) z(Y)) = E(w(X)) E(z(Y))$, for arbitrary w, z
- $\phi_{X,Y}(t,s) = \phi_X(t) \phi_Y(s)$, for arbitrary t and s

$P(X \leq x) = F(x)$, $P(Y \leq y) = G(y)$ and $P(X \leq x, Y \leq y) = H(x, y)$;
 f, g and h are the density analogs of F, G and H ; and
 ϕ denotes the Fourier transform

Imprecise probability independence

- Random-set independence
- Epistemic independence
- Strong independence
- Repetition independence
- Others?

Which should be called ‘independence’?

Notation

- X and Y are random variables
- F_X and F_Y are their probability distributions
- F_X and F_Y aren't known precisely, but we know they're within classes \mathcal{M}_X and \mathcal{M}_Y

$$X \sim F_X \in \mathcal{M}_X$$

$$Y \sim F_Y \in \mathcal{M}_Y$$

Strong independence

- $X \sim F_X \in \mathcal{M}_X$ and $Y \sim F_Y \in \mathcal{M}_Y$
- X and Y are stochastically independent
- All possible combinations of distributions from \mathcal{M}_X and \mathcal{M}_Y are allowed

$\Rightarrow X$ and Y are *strongly independent*

Complete absence of any relationship between X , Y

$$\mathcal{M}_{X,Y} = \{H : H(x, y) = F_X(x) F_Y(y), \\ F_X \in \mathcal{M}_X, F_Y \in \mathcal{M}_Y\}$$

Epistemic independence

- $X \sim F_X \in \mathcal{M}_X$ and $Y \sim F_Y \in \mathcal{M}_Y$
 - $\underline{E}(f(X)|Y) = \underline{E}(f(X))$ and $\underline{E}(f(Y)|X) = \underline{E}(f(Y))$ for all functions f
where \underline{E} is the smallest mean over all possible probability distributions
- $\Rightarrow X$ and Y are *epistemically independent*

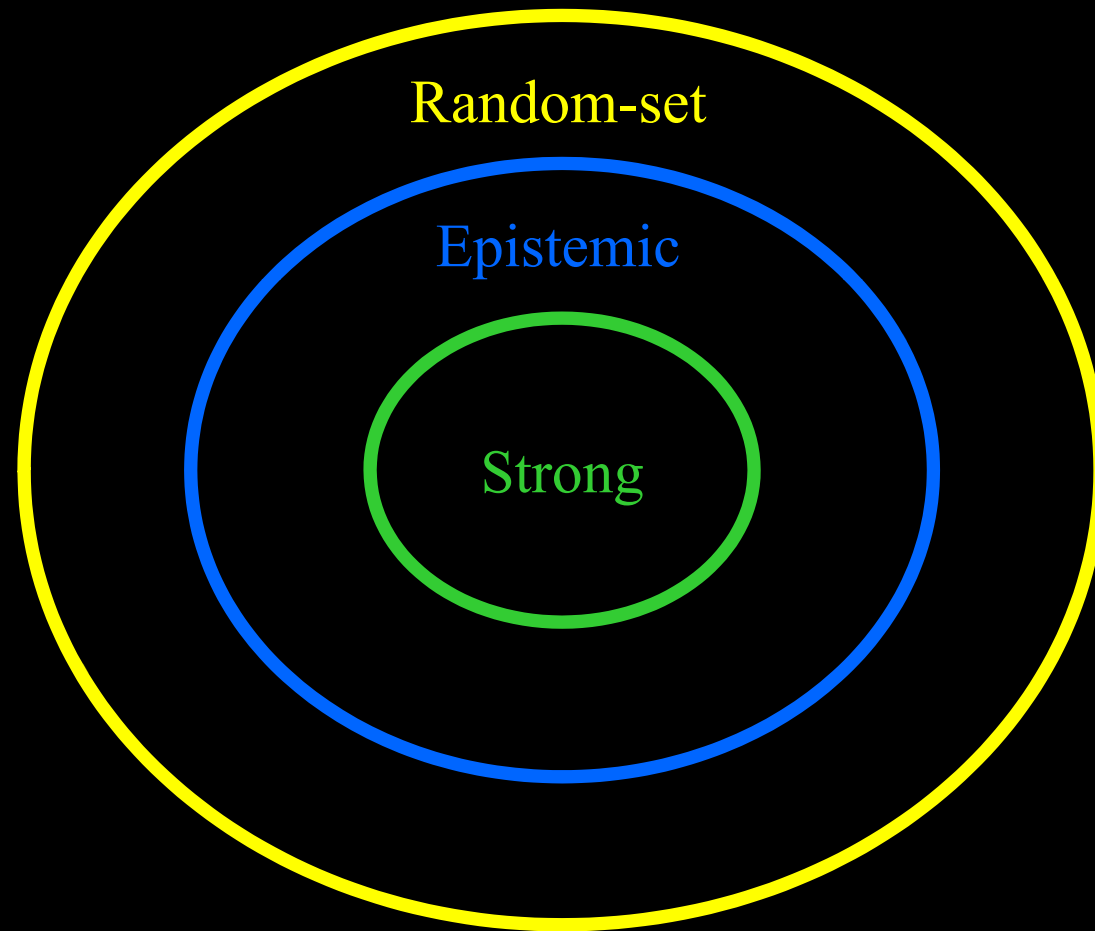
Lower bounds on expectations generalize the conditions $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$

Random-set independence

- Embodied in Cartesian products
- X and Y with mass functions m_X and m_Y are *random-set independent* if the Dempster-Shafer structure for their joint distribution has mass function $m(A_1 \times A_2) = m_X(A_1) m_Y(A_2)$ whenever A_1 is a focal element of X and A_2 is a focal element of Y , and $m(A) = 0$ otherwise
- Often easiest to compute

These cases of
independence
are *nested*.

Unknown = Fréchet



These cases of
independence
are *nested*.

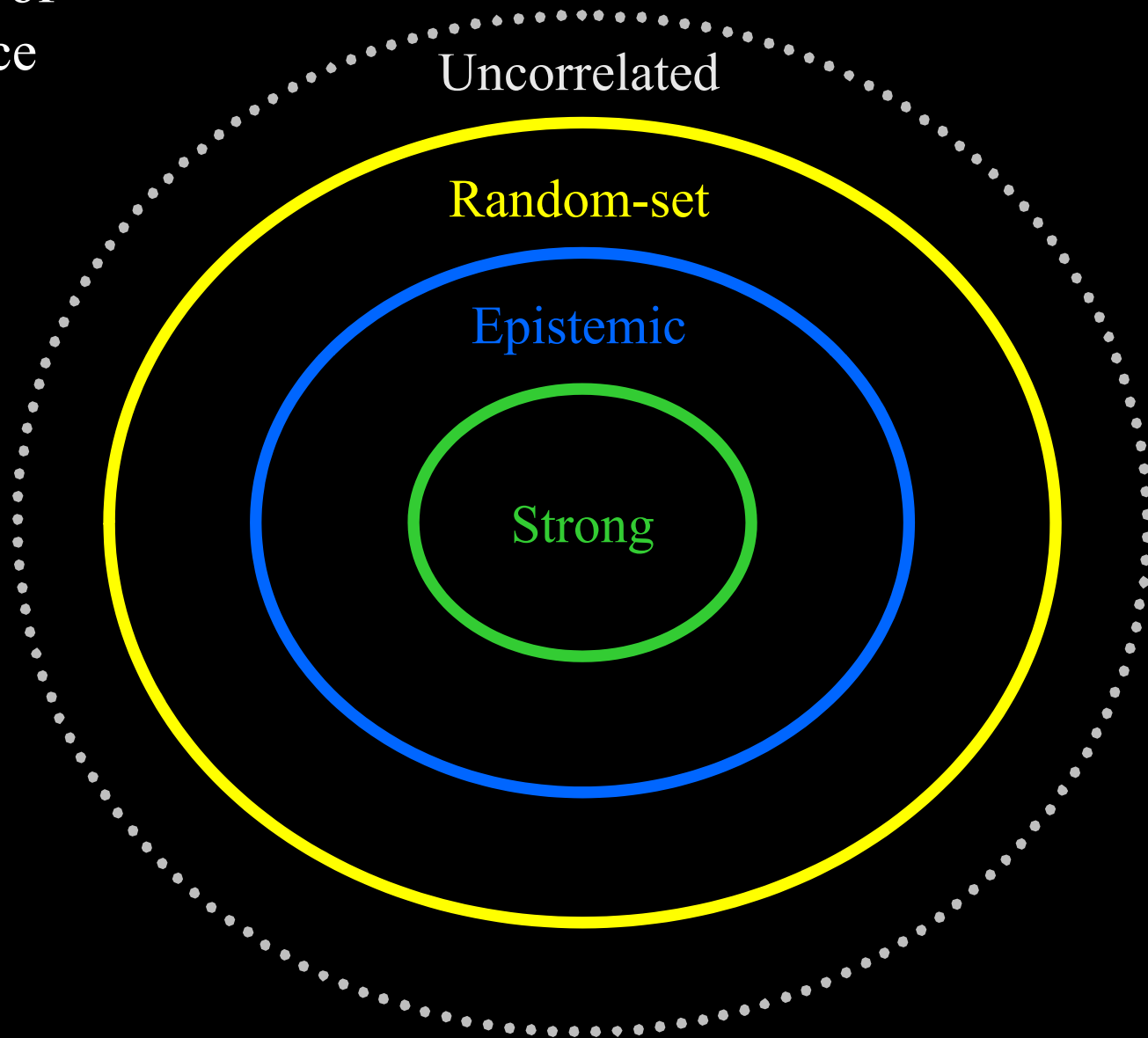
Unknown = Fréchet

Uncorrelated

Random-set

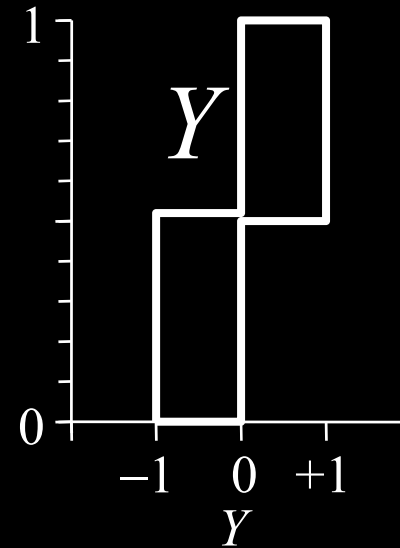
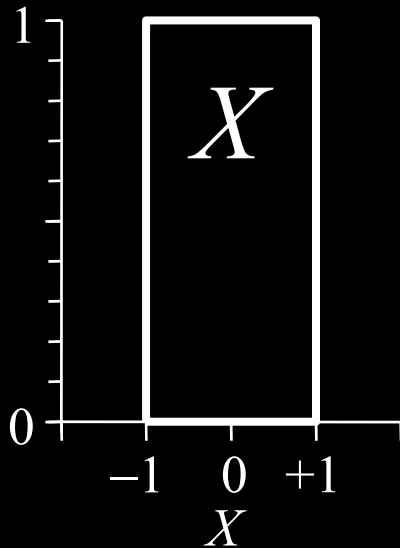
Epistemic

Strong



Interesting example

- $X = [-1, +1]$, $Y = \{([-1, 0], \frac{1}{2}), ([0, 1], \frac{1}{2})\}$

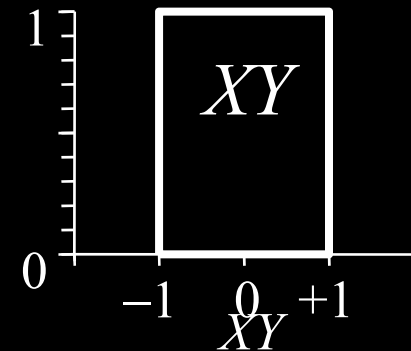


- If X and Y are “independent”, what is $Z = XY$?

Compute via Yager's convolution

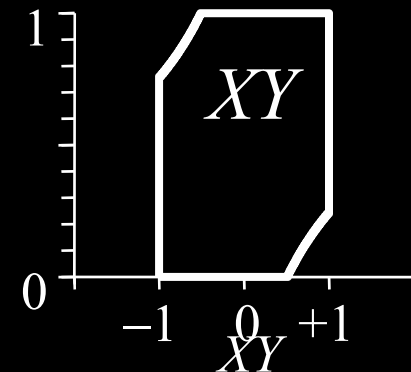
	Y	
	$([-1, 0], \frac{1}{2})$	$([0, 1], \frac{1}{2})$
X $([-1, +1], 1)$	$([-1, +1], \frac{1}{2})$	$([-1, +1], \frac{1}{2})$

The Cartesian product with one row and two columns produces this p-box



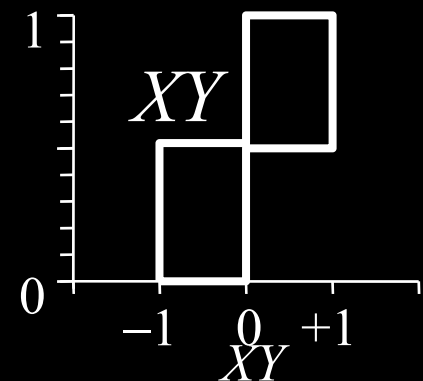
But consider the means

- Clearly, $EX = [-1, +1]$ and $EY = [-1/2, +1/2]$.
- Therefore, $E(XY) = [-1/2, +1/2]$.
- But if this is the mean of the product, and its range is $[-1, +1]$, then we know better bounds on the CDF.



And consider the quantity signs

- What's the probability P_Z that $Z < 0$?
- $Z < 0$ only if $X < 0$ or $Y < 0$ (but not both)
- $P_Z = P_X(1 - P_Y) + P_Y(1 - P_X)$, where
$$P_X = P(X < 0), \quad P_Y = P(Y < 0)$$
- But P_Y is $\frac{1}{2}$ by construction
- So $P_Z = \frac{1}{2}P_X + \frac{1}{2}(1 - P_X) = \frac{1}{2}$
- Thus, zero is the median of Z
- Knowing median and range improves bounds

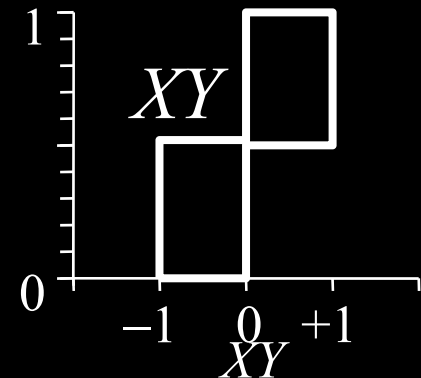
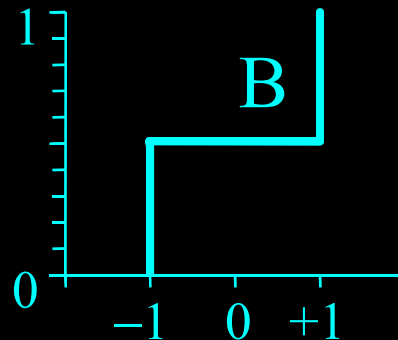


Best possible

- These bounds are realized by solutions

If $X = 0$, then $Z=0$

If $X = Y = B = \{(-1, \frac{1}{2}), (+1, \frac{1}{2})\}$, then $Z = B$



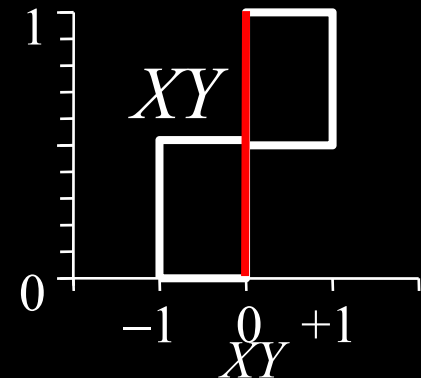
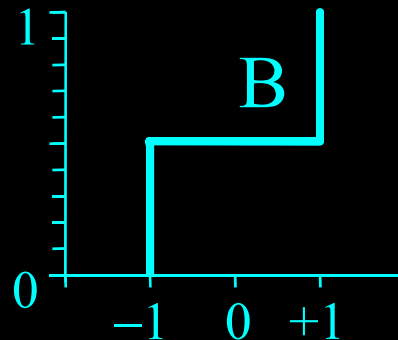
- So these bounds are also best possible

Best possible

- These bounds are realized by solutions

If $X = 0$, then $Z=0$

If $X = Y = B = \{(-1, \frac{1}{2}), (+1, \frac{1}{2})\}$, then $Z = B$



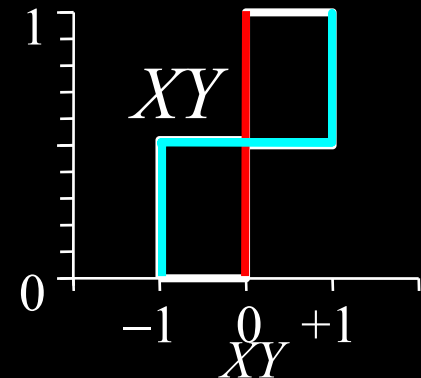
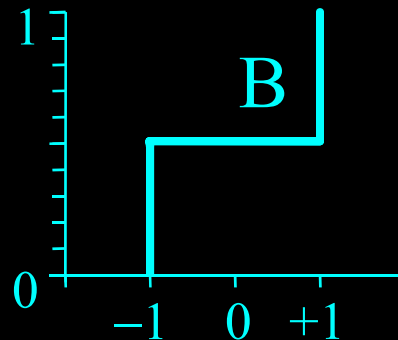
- So these bounds are also best possible

Best possible

- These bounds are realized by solutions

If $X = 0$, then $Z=0$

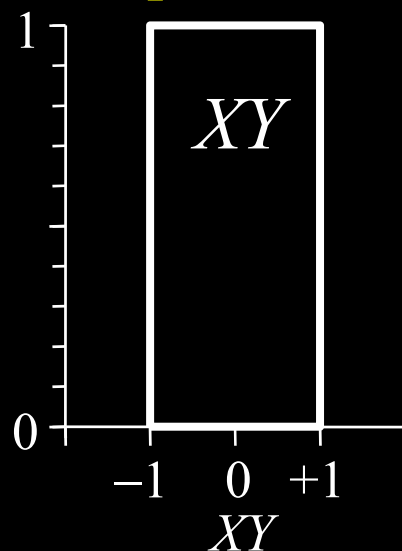
If $X = Y = B = \{(-1, \frac{1}{2}), (+1, \frac{1}{2})\}$, then $Z = B$



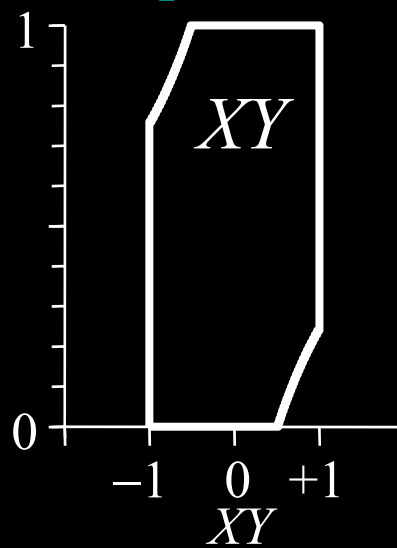
- So these bounds are also best possible

So which is correct?

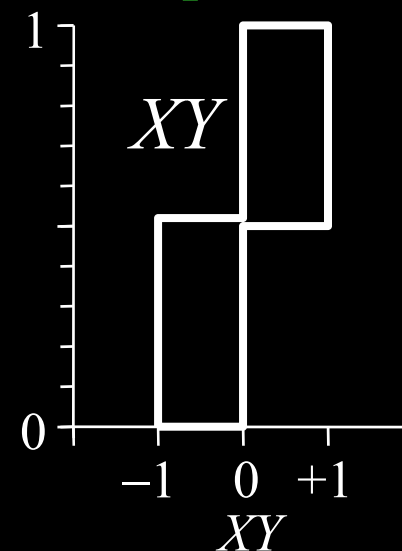
Random-set
independence



Moment
independence



Strong
independence



The answer depends on what one meant by “independent”.

So what?

- The example illustrates a practical difference between random-set independence and strong independence
- It disproves the conjecture that the convolution of uncertain numbers is not affected by dependence assumptions if at least one of them is an interval
- It shows convolutions with probability boxes and Dempster-Shafer structures may not be best-possible

Strategy for risk analysts

- Random-set independence is conservative
- Using the Cartesian product approach is always rigorous, though may not be optimal
- Convenient methods to obtain tighter bounds under other kinds of independence await derivation

Conclusions

Uncertain numbers

- P-boxes are very crude, but they can express the two main forms of uncertainty
- Despite their limitations, p-boxes may be useful for modeling uncertain numbers
- Simple arithmetic and logical expressions are easy to compute and understand

What p-boxes can't do

- Give best-possible bounds on non-tail risks
- Conveniently get best-possible bounds when dependencies are subtle
- Show what's most likely within the box

2MC simulations don't fill p-boxes

- 2-D Monte Carlo is not comprehensive
 - Inadequate model of ignorance
 - Dependence among parameters of a distribution
 - Uncertainty about dependence (Fréchet)
 - Non-denumerable model uncertainty
- Probability bounds analysis is not optimal
 - Independence between parameters of a distribution
 - Ternary (and higher) Fréchet operations

Maturing methodology

- Arithmetic
- Logical computations (and, or, not)
- Backcalculation, updating, deconvolution
- Decision analysis
- Statistics of data with interval uncertainty
- Sensitivity analysis
- Validation
- Non-linear ordinary differential equations
- Black-box strategies (Cauchy, quadratic, etc.)

Slide shows and/or papers on these topics are available on request

Web-accessible reading

<http://www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf>

(introduction to p-boxes and related structures)

<http://www.ramas.com/depend.zip>

(handling dependencies in probabilistic uncertainty modeling)

<http://www.ramas.com/bayes.pdf>

(introduction to Bayesian and robust Bayesian methods in risk analysis)

<http://www.ramas.com/intstats.pdf>

(statistics for data that may contain interval uncertainty)

<http://maths.dur.ac.uk/~dma31jm/durham-intro.pdf>

(Gert de Cooman's gentle introduction to imprecise probabilities)

<http://www.cs.cmu.edu/~qbayes/Tutorial/quasi-bayesian.html>

(Fabio Cozman's introduction to imprecise probabilities)

<http://idsia.ch/~zaffalon/events/school2004/school.htm>

(notes from a week-long summer school on imprecise probabilities)

Software

- Dan Berleant
 - Statool (free)
- Applied Biomathematics
 - PBDemo (free)
 - Risk Calc (commercial)
 - S3 and S4 packages for R (request beta version)

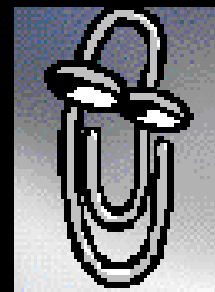
End

Ignorance and variability

- Bayesian approaches don't distinguish ignorance from equiprobability
- Neuroimaging and clinical psychology shows humans strongly distinguish uncertainty from risk
 - Most humans regularly and strongly deviate from Bayes
 - Hsu (2005) reported that people who have brain lesions associated with the site believed to handle uncertainty behave according to the Bayesian normative rules

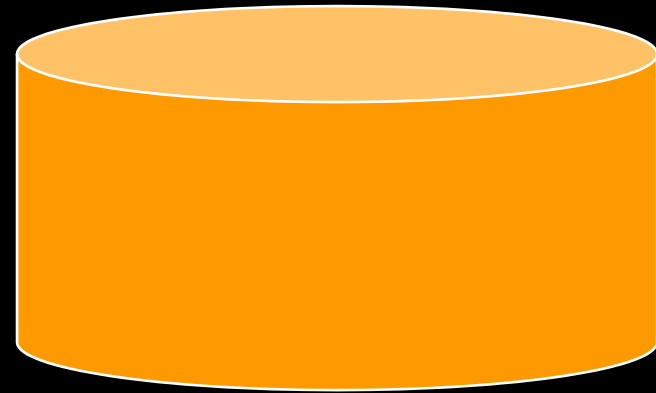
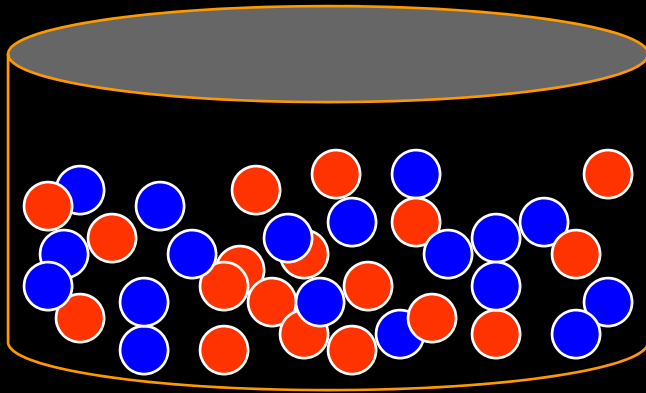
Ignorance and variability

- Bayesian approaches don't distinguish ignorance from equiprobability
- Neuroimaging and clinical psychology shows humans strongly distinguish uncertainty from risk
 - Most humans regularly and strongly deviate from Bayes
 - Hsu (2005) reported that people who have brain lesions associated with the site believed to handle uncertainty behave according to the Bayesian normative rules
- Bayesians are too sure of themselves (e.g.,

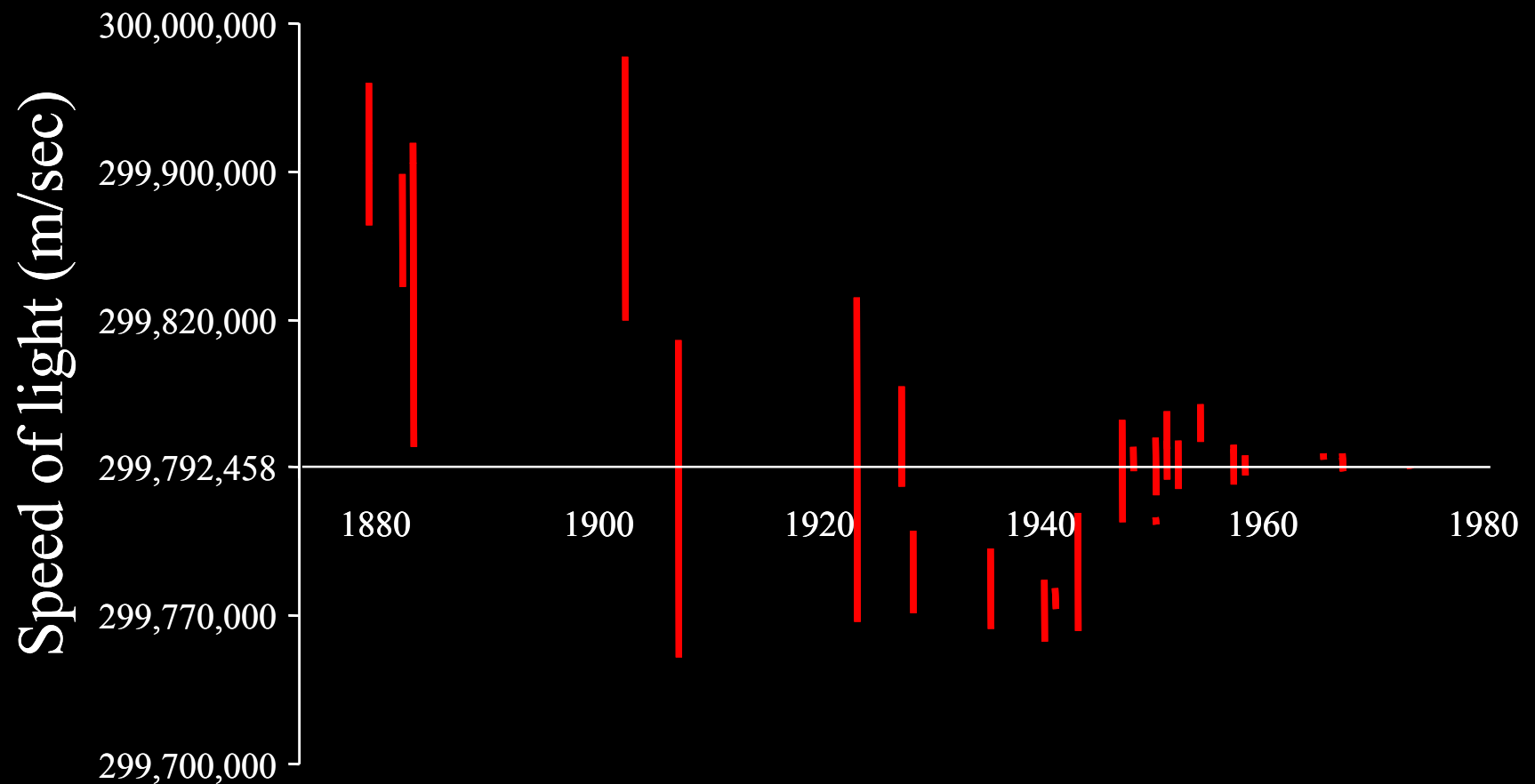


Variability v. uncertainty

- Precautionary principle
- Ellsberg paradox



History of the speed of light



History of overconfidence

- About 70% should enclose true value (fewer than half do)
- Overconfidence is “almost universal in all measurements of physical quantities” (Morgan and Henrion 1990)
- Humans (expert and otherwise) routinely grossly overconfident
90% confidence intervals typically enclose their true values only about 30 to 50% of the time
- Schlyakhter suggested we automatically widen all bounds

Everyone makes assumptions

- But not all sets of assumptions are equal!

Point value

Interval range

Entire real line

Normal distribution

Unimodal distribution

Any distribution

Linear function

Monotonic function

Any function

Independence

Known correlation

Any dependence

- Like to discharge unwarranted assumptions
“Certainties lead to doubt; doubts lead to certainty”

Two paths

- What assumptions are needed to get an answer?

It's always possible to find some

- What's the quantitative answer that doesn't depend on any unjustified assumptions?

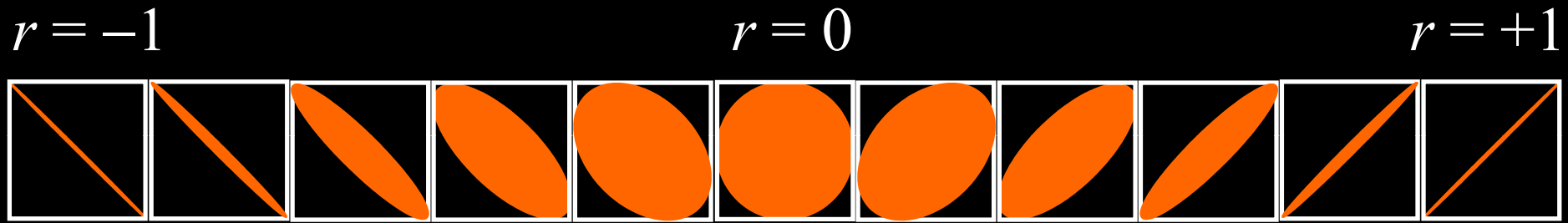
Recognizing when you've made an unjustified assumption may take some discipline

Sometimes, “I don't know” is the right answer

How to specify p-box inputs

- Sample data
when there's a lot, converges to a precise distribution
- Moment information
mean, median, mode, variance, range, etc.
- Structural information
unimodality, symmetry, positivity, (log)normality, etc.
- Modeling and allometry
express the problem in terms of subproblems
- Can also use precise distributions
fitted or assumed

Elliptic dependence



- Not complete (because $r=0$ isn't nondependence)

$$A+B = [-d, d] + (a_1 + a_2 + b_1 + b_2) / 2$$

$$d = \sqrt{(d_1^2 + d_2^2 + r d_1 d_2)}, d_1 = (a_2 - a_1)/2, d_2 = (b_2 - b_1)/2$$

Backcalculation

Backcalculation

- Needed for cleanup and remediation planning
- Untangles an equation in uncertain numbers when we know all but one of the variables
- For instance, backcalculation finds B such that $A+B = C$, from estimates for A and C

Hard with probability distributions

- Inverting the equation doesn't work
- Available analytical algorithms are unstable for almost all problems
- Except in a few special cases, Monte Carlo simulation cannot compute backcalculations; trial and error methods are required

Can't just invert the equation

prescribed

unknown

known


$$Dose = Concentration \times Intake$$

$$Concentration = Dose / Intake$$

When **concentration** is put back into the forward equation, the resulting **dose** is wider than planned

How come?

- Suppose dose should be less than 32, and intake ranges between 2 and 8
- If we solved for concentration by division, we'd get a distribution ranging between zero and 16
- But if we put that answer back into the equation

$$\text{Dose} = \text{Concentration} \times \text{Intake}$$

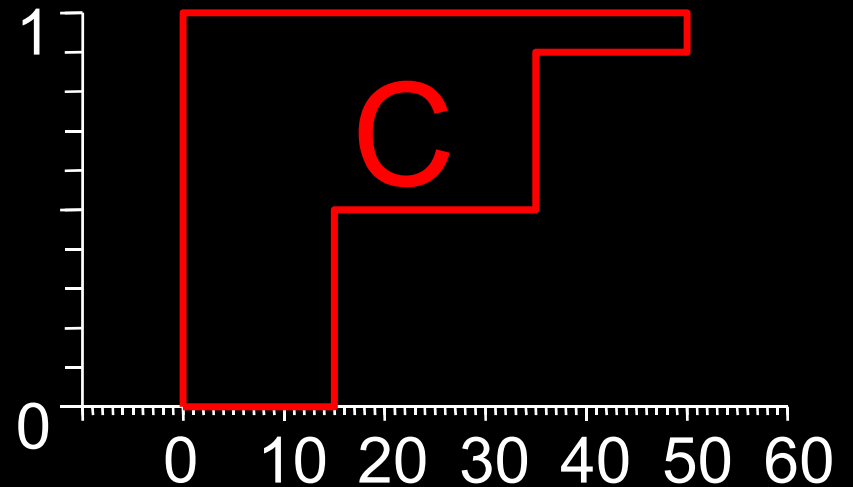
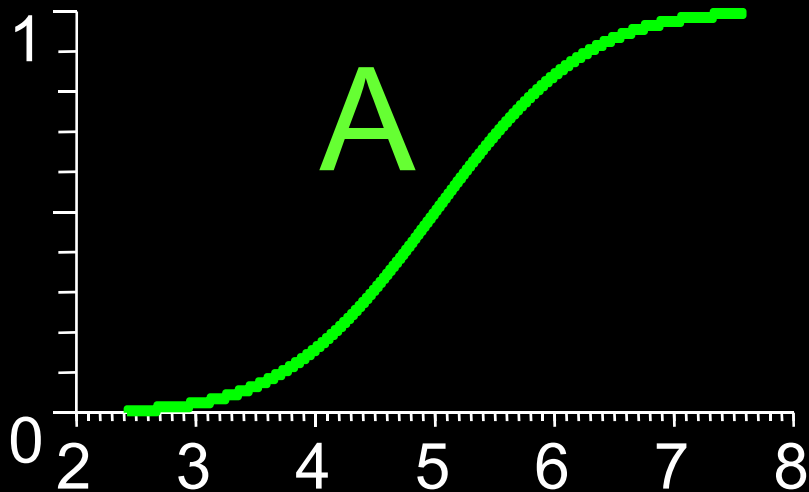
we'd get a distribution with values as large as 128, which is *four times larger than planned*

Backcalculation with p-boxes

Suppose $A + B = C$, where

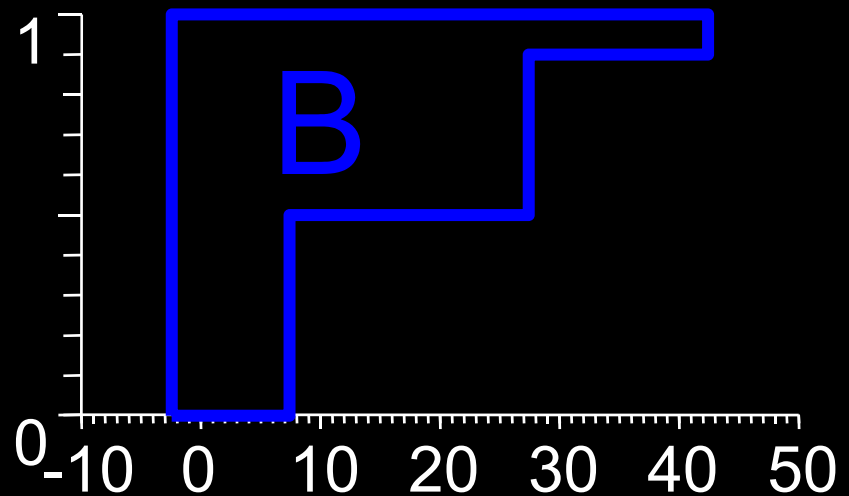
$A = \text{normal}(5, 1)$

$C = \{0 \leq C, \text{median} \leq 1.5, 90^{\text{th}} \text{ \%ile} \leq 35, \text{max} \leq 50\}$



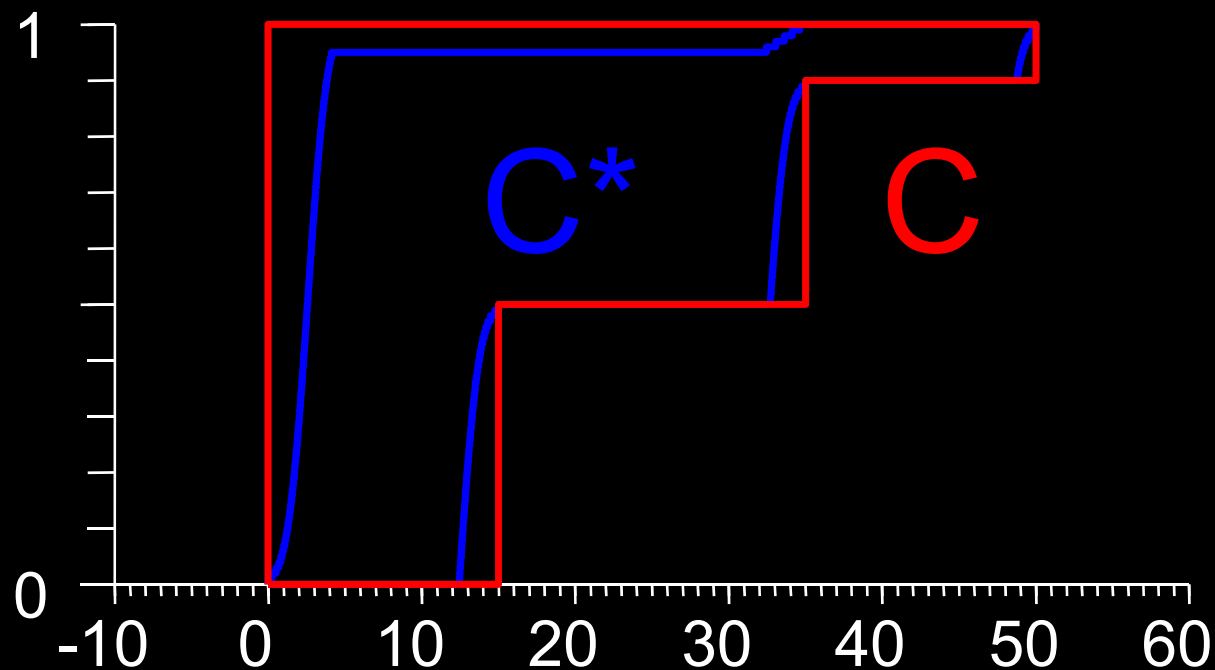
Getting the answer

- The backcalculation algorithm basically reverses the forward convolution
- Not hard at all...but a little messy to show
- Any distribution totally inside B is sure to satisfy the constraint ... it's a “kernel”



Check it by plugging it back in

$$A + B = C^* \subseteq C$$



Precise distributions don't work

- Precise distributions can't express the target
- A concentration distribution giving a prescribed distribution of doses seems to say we *want* some doses to be high
- Any distribution to the left would be better
- A p-box on the dose target expresses this idea

Conclusion

- Planning cleanup requires backcalculation
- Monte Carlo methods don't generally work except in a trial-and-error approach
- Can express the dose target as a p-box