

ECE 5/7383 INTRODUCTION TO QUANTUM INFORMATICS

Background Review

Estimators, risk, Fisher information, and the Cramér-Rao lower bound

Mitchell A. Thornton

Darwin Deason Institute for Cyber Security

Department of Electrical and Computer Engineering

Southern Methodist University

What this review covers

Part A fixes the vocabulary and the probability background. Parts B, C and D are the estimation theory, fifteen ideas in all, each applied to one example that runs through the whole deck.

A

Vocabulary and background

Notation glossary · random variables, PMF and PDF · the expectation operator · parameterized models · frequentist and Bayesian · likelihood · prior and posterior

B

Grading an estimator

1 Estimator · 2 Estimation error · 3 Loss function · 4 Risk · 5 Mean squared error

C

Comparing estimators

6 Risk as a function of the parameter · 7 Pointwise risk and dominance · 8 Bayes risk · 9 Minimax · 10 Sufficiency and Rao-Blackwell

D

Limits on any estimator

11 Consistency · 12 Score function · 13 Fisher information · 14 Cramér-Rao lower bound · 15 Maximum likelihood

Notation glossary: symbols

Every symbol used in this deck. A capital letter is a random variable, the matching lower case letter is a value it took.

Symbol	Read as	What it means
θ	<i>theta</i>	The unknown parameter. One fixed number we are trying to learn.
Θ	<i>capital theta</i>	The parameter space, the set of values the parameter could take.
X, X_i	<i>capital X</i>	A random variable, or the data before they are observed.
x, x_i	<i>lower case x</i>	A particular value the random variable took, a number in hand.
$\hat{\theta}$	<i>theta hat</i>	An estimator. The caret always marks a quantity computed from data.
$\tilde{\theta}$	<i>theta tilde</i>	A second estimator, usually an improved version of the first.
$\bar{X}, \bar{x}$	<i>X bar</i>	A sample mean. The bar always marks an average over the sample.
$f(x; \theta)$	<i>f of x given theta</i>	The density or mass function of the data at parameter value theta.
$X \sim N(\theta, \sigma^2)$	<i>X is distributed as</i>	Says which distribution the random variable follows.
$\propto$	<i>proportional to</i>	Equal up to a factor that does not depend on the parameter.

Notation glossary: operators

These are the operators that appear without being defined in most textbooks.

Symbol	Read as	What it means
$E[\cdot]$	<i>expectation of</i>	The long-run average value, defined formally three slides on.
$\text{Var}(\cdot)$	<i>variance of</i>	The average squared distance from the mean, a measure of spread.
$\min_{\theta} g(\theta)$	<i>the minimum over theta</i>	The smallest value the function takes. Returns a value.
$\arg \min_{\theta} g(\theta)$	<i>the argmin over theta</i>	The theta at which that smallest value occurs. Returns a location.
$\sup_{\theta} g(\theta)$	<i>the supremum over theta</i>	The least upper bound. Used instead of max when no largest value is attained.
$\sum_{i=1}^n$	<i>sum over i</i>	Add up the terms. The discrete counterpart of an integral.
$\int_{-\infty}^{\infty}$	<i>integral over x</i>	The continuous counterpart of a sum.
$\mathbf{1}\{A\}$	<i>indicator of A</i>	Equals 1 when A is true and 0 when it is false.
$f(x \mid \theta)$	<i>f of x given theta</i>	A conditional density. The bar separates what is random from what is held fixed.
$\ln, \ell(\theta)$	<i>natural log, script ell</i>	Natural logarithm, and the log-likelihood function it produces.

The words, in plain English

Each term is defined properly later. This slide is the one-line version to keep in mind while reading.

Describing the data

Random variable	A numerical outcome that is not known before the experiment.
Statistic	Any number computed from the data alone, with no unknowns in it.
Parameter	A fixed unknown number that indexes which distribution generated the data.
Estimator	A rule that turns data into a guess for the parameter.
Estimate	The number that rule returns on one particular data set.
Likelihood	The density of the data already observed, read as a function of the parameter.

Grading the guess

Loss	The cost charged for one wrong guess.
Risk	The average loss over every data set the model could have produced.
Bias	How far the estimator is off on average.
Prior	What we believed about the parameter before seeing any data.
Posterior	What we believe about it afterwards.
Efficiency	How close an unbiased estimator comes to the smallest variance possible.

Random variables, PMF and PDF

A random variable assigns a number to the outcome of an experiment. Its distribution says how that number is spread out, and takes one of two forms depending on whether the values are countable.

Discrete

probability mass function

$$p(x) = P(X = x), \quad \sum_x p(x) = 1$$

Each value carries a chunk of probability. The mass function is a probability, so it never exceeds one.

$$P(X = x) = p(x)$$

Examples: A coin, a die, a count of successes in n trials.

Continuous

probability density function

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

Probability is spread over a continuum, so it attaches to intervals rather than to points. The density is not a probability and may exceed one.

$$P(X = x) = 0$$

Examples: A weight, a voltage, a waiting time.

Shared by both

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

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

The cumulative distribution function is defined the same way in both cases, which is why it is the safer object to reason with.

The expectation operator

Expectation is the probability-weighted average of a quantity. It is the operator every definition in this deck is built on, since risk is an expectation of loss.

Discrete

$$E[g(X)] = \sum_x g(x) p(x)$$

Weight each value by its probability and add.

Continuous

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Weight each value by its density and integrate.

the same idea twice, written for the two kinds of distribution

Properties used throughout this deck

$$E[aX + bY + c] = a E[X] + b E[Y] + c$$

Linearity. Expectation passes through sums and constants, whatever the dependence between the variables.

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

Variance written as an expectation, which is how the bias-variance split is derived.

$$E[X] = E[E[X | Y]]$$

The tower property. Averaging a conditional average recovers the overall average, and this is what makes Rao-Blackwell work.

SUBSCRIPT CONVENTION

A subscript on the expectation names the parameter value the averaging is done at. Different parameter values give different averages.

$$E_{\theta}[\bar{X}] = \theta, \quad \text{Var}_{\theta}(\bar{X}) = \frac{\sigma^2}{n} = 9 \text{ g}^2$$

Both are computed with the parameter held fixed.

A parameterized statistical model

A model is not one distribution. It is a whole family of them, indexed by a parameter. Estimation is the job of deciding which member of the family produced the data.

$$\{ f(x; \theta) : \theta \in \Theta \}$$

read as: the set of densities f , one for each value of θ in the parameter space

x is the VARIABLE

What we observe. It is random, it changes from one data set to the next, and it ranges over the sample space.

θ is the PARAMETER

What we do not know. It is one fixed number, it does not change with the data, and it ranges over the parameter space.

The semicolon in $f(x; \theta)$ is deliberate. It separates the variable from the parameter and signals that θ is not being treated as random. A vertical bar would say the opposite, and that is exactly the switch the Bayesian approach makes.

The model used in this deck

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \theta)^2}{2\sigma^2}\right)$$

NUISANCE PARAMETERS

The noise variance appears in the formula but is assumed known, so it is not being estimated. An unknown quantity we must carry but do not care about is called a nuisance parameter.

The running example

One statistical model is used for every definition in this deck, and one data set is used for every number.

$$X_1, X_2, \dots, X_n \text{ i.i.d. } N(\theta, \sigma^2), \quad \sigma \text{ known, } \theta \text{ unknown}$$

A crate of apples has an unknown mean weight. Each weighing on the scale returns that mean plus independent measurement noise, so one weighing is one draw from a normal distribution whose mean is the quantity we want.

What is known, and what is not

the shape of the noise distribution	normal
the noise standard deviation	15 g
the number of weighings	25
the mean weight	unknown, this is θ

THE DATA SET

The 25 weighings used from here on average to 148.2 g. That one number is the input to every worked example in the deck.

sample mean	148.2 g
true mean	150 g (we do not see this)
noise variance	225 g²
standard error	3 g

The true value is written down only so the arithmetic can be checked.

Frequentist and Bayesian approaches

The two traditions disagree about one thing: whether it makes sense to put a probability distribution on the parameter. Everything else follows from that.

Frequentist

θ is fixed, the data are random

What is random	The data. The parameter is one fixed unknown number.
What probability describes	The behaviour of the procedure over repeated samples.
What you need up front	Only the model. No prior is required.
What you report	An estimate, a standard error, and a confidence interval.
How to read the interval	The interval is random. Over repeated samples, 95 percent of such intervals cover θ .

Bayesian

θ is random, the data are what we condition on

What is random	The parameter, as a way of expressing what we do not know.
What probability describes	Our degree of belief, updated by the data.
What you need up front	The model and a prior distribution on the parameter.
What you report	A posterior distribution, often summarized by its mean and spread.
How to read the interval	Given the data and the prior, θ lies in the interval with 95 percent probability.

Both use the same likelihood, and this deck uses both. Pointwise risk is a frequentist criterion, Bayes risk averages that same risk against a prior, and the two meet on the minimax slide.

Likelihood and log-likelihood

The likelihood is the same formula as the density, read the other way round. The density treats the parameter as fixed and the data as variable. The likelihood fixes the data that were actually observed and lets the parameter vary.

$$L(\theta; x) = f(x; \theta)$$

same formula, different argument

For independent observations the likelihood multiplies

$$L(\theta; x) = \prod_{i=1}^n f(x_i; \theta)$$

Taking logs turns the product into a sum

$$\ell(\theta) = \ln L(\theta; x) = \sum_{i=1}^n \ln f(x_i; \theta)$$

- The logarithm is strictly increasing, so it does not move the maximizer. Maximizing either one gives the same estimator.
- Sums differentiate term by term and do not underflow the way a product of many small numbers does.
- The score and the Fisher information are both built from the log-likelihood, which is why it appears throughout the second half of this deck.

A COMMON MISTAKE

The likelihood is not a probability distribution over the parameter.

$$\int L(\theta; x) d\theta \neq 1$$

RUNNING EXAMPLE

The log-likelihood for the apple model.

$$\ell(\theta) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2$$

Only the final sum involves the parameter.

Prior, likelihood, and posterior

A prior is a probability distribution on the parameter, written down before the data are seen. Bayes' rule combines it with the likelihood to give the posterior, which is what we believe afterwards.

$$\pi(\theta \mid x) = \frac{f(x \mid \theta) \pi(\theta)}{\int f(x \mid \theta') \pi(\theta') d\theta'}$$

the denominator does not involve theta, so it is only a normalizing constant

Prior

What is believed before the data. It can encode past experiments, physical limits, or deliberate vagueness.

Likelihood

What the data say, through the model. This is the only place the observations enter.

Posterior

The two combined. It is a full distribution, not a single number, and it is the Bayesian answer.

$$\pi(\theta \mid x) \propto f(x \mid \theta) \pi(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

To turn the posterior into a single estimate, summarize it. **Under squared error loss the best summary is the posterior mean**, which is the link back to the Bayes risk slide later in the deck.

$$\hat{\theta}_B = E[\theta \mid x]$$

Worked example: the frequentist approach

The 25 weighings, with the parameter treated as one fixed unknown number.

The estimate

$$\hat{\theta} = \bar{x} = 148.2 \text{ g}$$

Its standard error

$$\text{se} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3 \text{ g}$$

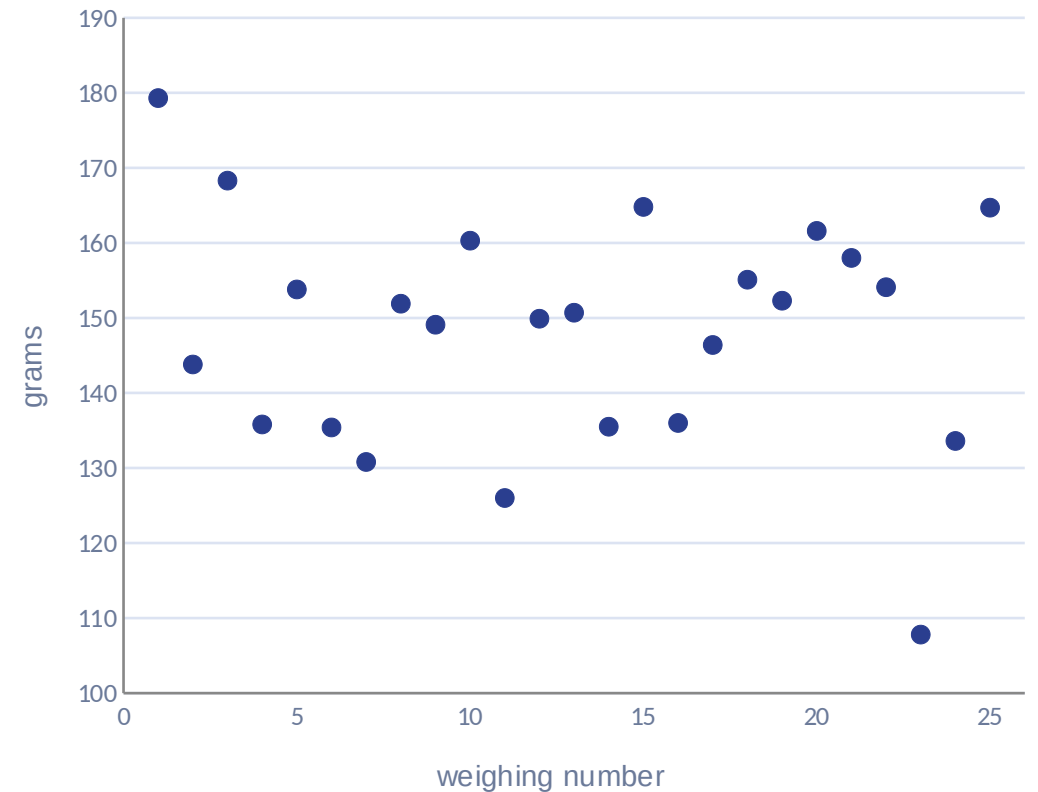
A 95 percent confidence interval

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = [142.3, 154.1]$$

HOW TO READ THE INTERVAL

The parameter is not random, so it is wrong to say it lies in this interval with probability 0.95. The interval is the random thing. If the whole experiment were repeated many times, 95 percent of the intervals built this way would contain the true mean.

The 25 weighings, averaging 148.2 g



Each reading scatters around the unknown mean. Averaging them shrinks the scatter by the square root of 25.

Worked example: the Bayesian approach

The same 25 weighings, now with a prior on the mean weight. Suppose the grower claims the crate averages 155 g, and we hold that belief loosely.

The prior

$$\pi(\theta) = N(155, 10^2)$$

Normal prior and normal likelihood give a normal posterior

$$\theta | x \sim N(\mu_n, \tau_n^2)$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2} = \frac{1}{100} + \frac{1}{9} = \frac{109}{900}$$

$$\mu_n = w \bar{x} + (1 - w) \mu_0, \quad w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}$$

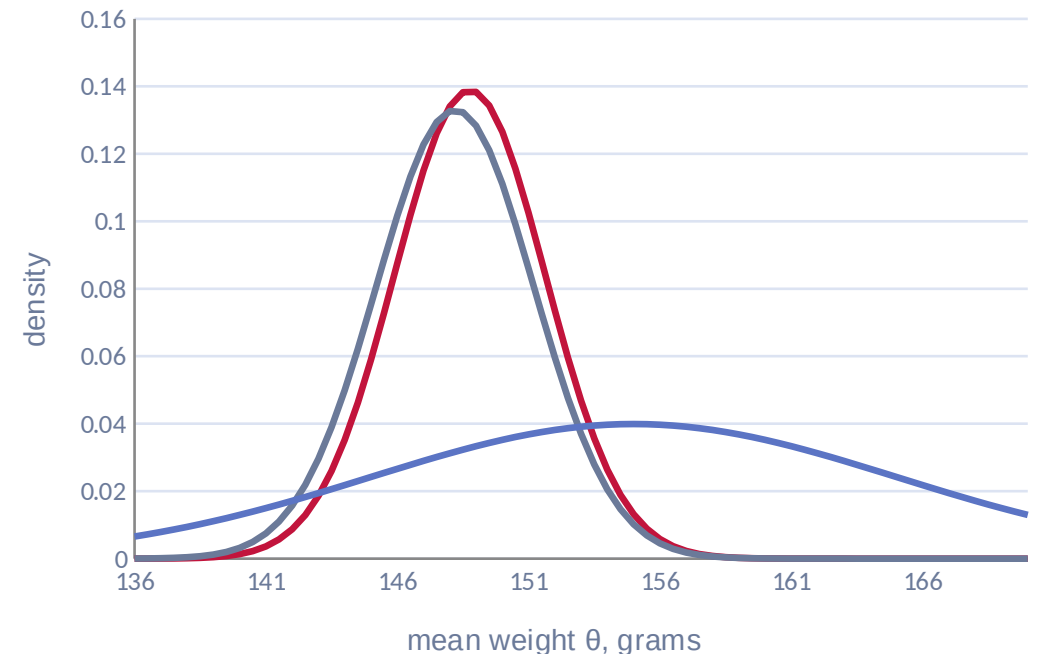
RESULT

$$w = \frac{100}{109} = 0.917, \quad \mu_n = 148.76 \text{ g}, \quad \tau_n = 2.87 \text{ g}$$

$$148.76 \pm 1.96 (2.87) = [143.1, 154.4]$$

The data carry 91.7 percent of the weight, so the posterior sits close to the sample mean, pulled 0.56 g toward the prior. Its spread is **narrower** than the standard error, because the prior adds information.

Prior, likelihood and posterior



— prior — likelihood, rescaled — posterior

$$\tau \rightarrow \infty \Rightarrow \mu_n \rightarrow \bar{x}$$

The likelihood is rescaled to fit the axis, since it is not a density in θ . A vaguer prior slides the posterior onto the likelihood, and the two approaches then agree.

1 Estimator

An estimator is a rule that assigns a numerical guess for the parameter to every possible data set. It is a function of the data alone, so it never contains the unknown parameter.

$$\hat{\theta}(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

the sample mean, our estimator throughout this deck

- The estimator is the rule. The number it returns on one data set is the estimate.
- Decision theory calls the same object a decision rule and often writes it as a lower case delta of the data.
- Many rules compete for the same parameter. The sample median and the average of the first half of the sample are also estimators of the mean.

RUNNING EXAMPLE

If the 25 weighings average 148.2 g, the estimate is 148.2 g. A different crate sample would give a different number from the same rule.

That sample to sample variation is what the next slide measures.

2

Estimation error has a random part and a systematic part

The error of an estimator on one data set is the difference between the guess and the truth.

$$e = \hat{\theta} - \theta$$

$$\hat{\theta} - \theta = \underbrace{(\hat{\theta} - E_{\theta}[\hat{\theta}])}_{\text{random part}} + \underbrace{(E_{\theta}[\hat{\theta}] - \theta)}_{\text{systematic part}}$$

random part

systematic part

The split is exact and holds for every estimator in every model.

Random error

Varies from one data set to the next and averages to zero. Its size is measured by the standard deviation of the estimator, also called the standard error.

Systematic error

The same on average for every data set. This is the bias, the gap between the mean of the estimator and the parameter.

$$b(\theta) = E_{\theta}[\hat{\theta}] - \theta$$

An estimator with zero bias at every parameter value is called unbiased.

RUNNING EXAMPLE

The sample mean is unbiased for the mean, so all of its error is random.

$$b(\theta) = 0, \quad \text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{15}{5} = 3 \text{ g}$$

A typical weighing average sits within a few grams of the true mean.

3 Loss function

A loss function assigns a cost to a single guess, measured against the parameter value that actually produced the data. It encodes what kind of mistake we care about.

$$L(\theta, a) = (a - \theta)^2$$

squared error loss, the choice used throughout this deck

Other common choices

$$L(\theta, a) = |a - \theta|$$

absolute error

$$L(\theta, a) = \mathbf{1}\{a \neq \theta\}$$

zero one loss

- Squared error is smooth, differentiable, and penalizes a large miss far more than a small one.
- Absolute error is less sensitive to rare large errors.
- Zero one loss charges a flat cost for any miss and suits discrete parameters.

RUNNING EXAMPLE

Suppose the true mean weight is 150 g and one data set leads us to guess 140 g.

$$L(150, 140) = (140 - 150)^2 = 100 \text{ g}^2$$

Loss is the cost of that one guess. It is not yet a property of the rule, because a different sample would give a different loss.

4

Risk is expected loss

Loss grades one guess. Risk grades the rule, by averaging the loss over every data set the model could produce with the parameter held fixed.

$$R(\theta, \hat{\theta}) = E_{\theta}[L(\theta, \hat{\theta}(X))]$$

the expectation is over the data, not over the parameter

- Risk is a property of the estimator and the model together, available before any data are collected.
- It cannot be computed from a single data set, since it averages over all of them.
- Comparing two estimators means comparing two risk functions, which is why the next slides look at how risk varies with the parameter.

RUNNING EXAMPLE

For the sample mean under squared error loss the risk has a closed form.

$$R(\theta, \bar{X}) = \frac{\sigma^2}{n} = \frac{225}{25} = 9 \text{ g}^2$$

On average the squared miss is 9 g^2 , whatever the true mean happens to be.

Squared error loss makes the risk a mean squared error

Substituting squared error loss into the definition of risk gives the familiar mean squared error.

$$R(\theta, \hat{\theta}) = E_{\theta}[(\hat{\theta} - \theta)^2] = \text{MSE}_{\theta}(\hat{\theta})$$

$$\text{MSE}_{\theta}(\hat{\theta}) = \text{Var}_{\theta}(\hat{\theta}) + [b(\theta)]^2$$

random part

systematic part

Take the exact error split from earlier, square it, and the cross term vanishes.

- Two estimators can have equal mean squared error with very different mixes of bias and variance.
- Allowing a little bias sometimes buys a large reduction in variance, which is the basis of shrinkage and regularization.
- For an unbiased estimator the mean squared error is exactly the variance, and that is the case the Cramér-Rao bound addresses.

RUNNING EXAMPLE

The sample mean is unbiased, so the squared bias term is zero.

$$\text{MSE} = \frac{\sigma^2}{n} + 0^2 = 9 \text{ g}^2 \quad \Rightarrow \quad \text{RMSE} = 3 \text{ g}$$

Reporting the root mean squared error of 3 g restores the original units.

6 Risk is a function of the parameter

Risk is evaluated with the parameter held fixed, so each parameter value has its own risk. In general the distribution of the data changes with the parameter, so the risk changes too.

Normal model, known noise variance

$$R(\theta, \bar{X}) = \frac{\sigma^2}{n}$$

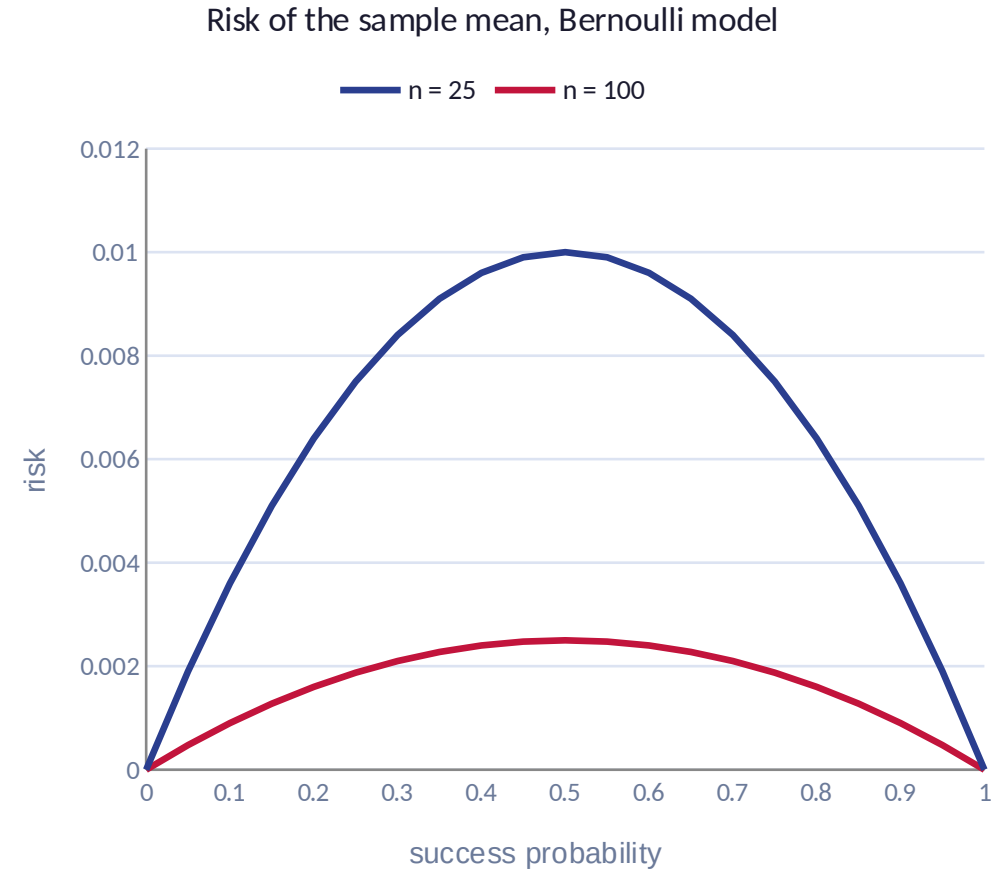
A convenient special case. The variance does not involve the mean, so the risk of the sample mean is flat in the parameter.

Bernoulli model, estimating the success probability

$$R(\theta, \bar{X}) = \frac{\theta(1-\theta)}{n}$$

Here the variance is tied to the parameter, so the same estimator is harder to pin down near one half than near either end. The curve on the right shows the effect.

Because a flat risk function is the exception rather than the rule, an estimator is compared against its rivals across the whole parameter range.



7 Pointwise risk, dominance, and admissibility

The pointwise risk is the risk read off at one chosen parameter value. It answers how well the rule performs if the truth happens to be that value.

RUNNING EXAMPLE

$$R(150, \bar{X}) = \frac{\sigma^2}{n} = 9 \text{ g}^2$$

One estimator dominates another when

$$R(\theta, \hat{\theta}_1) \leq R(\theta, \hat{\theta}_2) \quad \text{for all } \theta$$

- with strict inequality somewhere. The dominated rule is then inadmissible and should not be used.
- Pointwise comparison rarely settles anything on its own, because one estimator usually wins on part of the parameter range and loses on the rest.
- When no rule dominates, the whole risk function has to be collapsed into one number. The next two slides give the two standard ways.

WHY THIS MATTERS

Reporting a single risk number without saying where it was evaluated hides the comparison. Two estimators quoted at different parameter values cannot be ranked.

For the normal model used here the pointwise risk is the same at every parameter value, so the ambiguity does not arise. That is a feature of this model rather than a general rule.

Admissibility is a weak requirement. Many poor estimators are admissible, including the rule that ignores the data and always returns the same number.

8 Bayes risk

Bayes risk collapses the risk function to one number by averaging it against a prior weighting on the parameter.

$$r(\pi, \hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta$$

the prior says which parameter values deserve weight

What it buys

- One number per estimator, so any two rules can be ranked without ambiguity.
- The rule that minimizes it is the posterior mean under squared error loss.
- Every Bayes rule with finite Bayes risk is admissible.

What it costs

- The ranking depends on the prior, and a different prior can reverse it.
- A prior has to be chosen and defended before any comparison can be made.
- Minimax, on the next slide, avoids the choice at the price of pessimism.

RUNNING EXAMPLE

$$\pi: \theta \sim N(\mu_0, \tau^2) \quad r(\pi, \bar{X}) = \int \frac{\sigma^2}{n} \pi(\theta) d\theta = \frac{\sigma^2}{n}$$

flat risk makes the prior irrelevant here

The minimax criterion grades an estimator by the worst risk it ever incurs, then picks the rule whose worst case is least bad. No prior is required.

$$\hat{\theta}_M = \arg \min_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$$

the inner supremum is over the parameter, the outer minimum is over estimators

How to read it

- It is a guarantee. Whatever the parameter turns out to be, the risk is at most this number.
- It ignores how likely each parameter value is, so it can be pessimistic.

How one is found

A rule whose risk is the same at every parameter value, and which is Bayes for some prior, is minimax. Constant risk is the usual signature.

$$R(\theta, \hat{\theta}_M) = \text{constant in } \theta$$

Relation to Bayes

- A minimax rule is the Bayes rule for the least favorable prior.
- Bayes optimizes the average case, minimax the worst case.
- They agree when the risk of the Bayes rule happens to be constant.

The normal model used elsewhere in this deck has flat risk, so it cannot illustrate the idea. The next slide works the Bernoulli case, where the sample proportion is not minimax.

Minimax worked example

The normal model has flat risk, so it cannot show the idea. This example uses the Bernoulli model with $n = 25$ trials, estimating the success probability under squared error loss.

The sample proportion

$$\sup_{\theta} R(\theta, \bar{X}) = \frac{1}{4n} = \frac{1}{100} = 0.0100$$

Unbiased, and its risk peaks at one half. The worst case is one in a hundred.

The minimax rule

$$\hat{\theta}_M = \frac{X + \sqrt{n}/2}{n + \sqrt{n}} \quad R(\theta, \hat{\theta}_M) = \frac{1}{4(1 + \sqrt{n})^2}$$

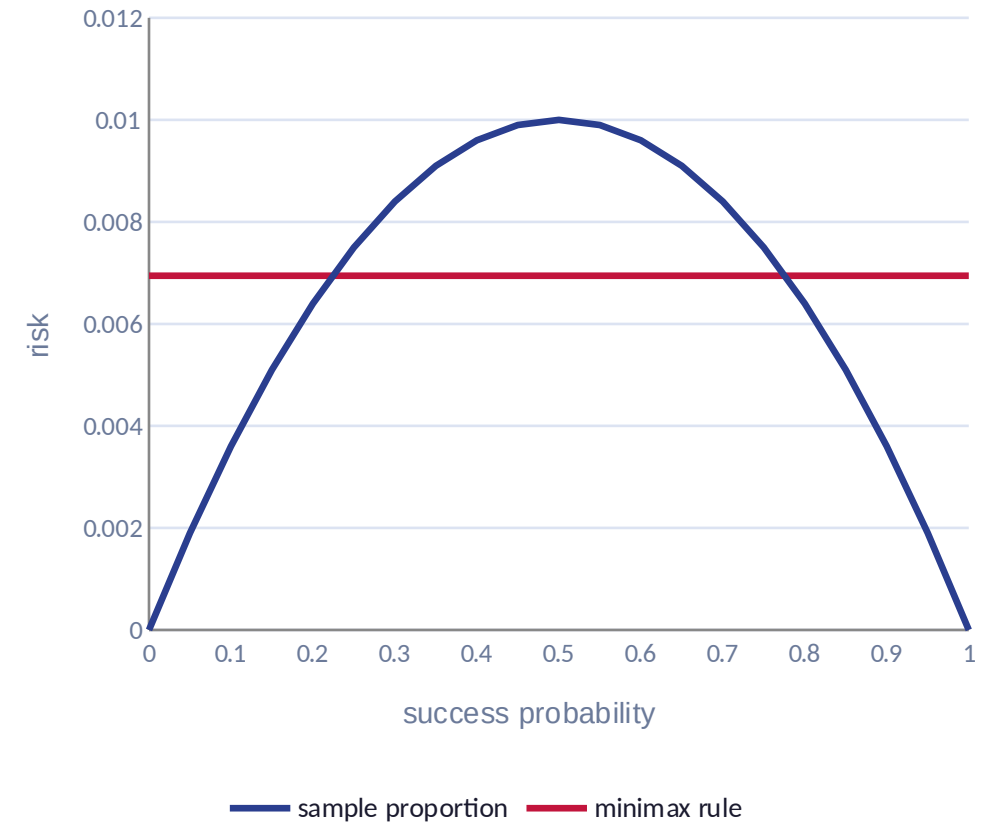
Slightly biased, pulling every estimate toward one half. Its risk does not depend on the parameter at all, which is the signature of a minimax rule.

AT $n = 25$

$$\hat{\theta}_M = \frac{X + 2.5}{30} \quad R(\theta, \hat{\theta}_M) = \frac{1}{144} = 0.00694$$

Worst case improves by a factor of **1.44**. The sample proportion is still better outside the range 0.224 to 0.776, where its risk falls below the constant.

Risk at $n = 25$, Bernoulli model



Sufficiency and the Rao-Blackwell theorem

A statistic is sufficient when it carries every piece of information the data hold about the parameter. The density then factors into a part that involves the parameter only through that statistic, and a part that does not involve it at all.

$$f(x; \theta) = g(T(x); \theta) h(x)$$

the factorization criterion

Conditioning on a sufficient statistic never hurts

$$\tilde{\theta} = E[\hat{\theta} | T(X)] \quad \text{Var}_{\theta}(\tilde{\theta}) \leq \text{Var}_{\theta}(\hat{\theta})$$

- The improved rule has the same bias, so an unbiased rule stays unbiased.
- The variance is never larger, and is strictly smaller unless the original rule was already a function of the sufficient statistic.
- This is why the search for a best unbiased estimator can be restricted to functions of a sufficient statistic.

RUNNING EXAMPLE

$$T(X) = \sum_{i=1}^n X_i \quad \text{is sufficient for } \theta$$

Start from a deliberately wasteful unbiased rule, the first weighing on its own, and condition it on the total.

$$E[X_1 | \bar{X}] = \bar{X} : \quad 225 \text{ g}^2 \longrightarrow 9 \text{ g}^2$$

The variance falls by a factor of 25, and the improved rule is the sample mean.

Consistency

Everything so far has held the sample size fixed. Consistency asks a different question: does the estimator settle on the right answer as data accumulate?

$$\hat{\theta}_n \longrightarrow \theta \quad \text{as } n \rightarrow \infty$$

convergence in probability, for every value the parameter might take

A sufficient condition that is easy to check

$$b_n(\theta) \rightarrow 0 \quad \text{and} \quad \text{Var}_\theta(\hat{\theta}_n) \rightarrow 0 \quad \Rightarrow \quad \text{consistent}$$

- Consistency is a large sample property. It says nothing about how the rule behaves at the sample size actually in hand.
- An estimator can be consistent and still badly biased at small n , and it can be unbiased at every n yet inconsistent.
- It is a minimum requirement rather than a recommendation.

RUNNING EXAMPLE

The sample mean is unbiased at every sample size and its variance falls as one over n , so it is consistent.

$$\text{sd}(\bar{X}) = \frac{15}{\sqrt{n}} \leq 1 \text{ g} \quad \Leftrightarrow \quad n \geq 225$$

Driving the standard error down to a single gram takes 225 weighings, nine times the 25 we have.

The score function

Hold the data fixed and ask how fast the log density changes as the parameter moves. That derivative is the score, and it measures how strongly the observed data prefer one parameter value over its neighbors.

$$s(\theta; x) = \frac{\partial}{\partial \theta} \ln f(x; \theta)$$

the derivative is in the parameter, with the data held fixed

Two properties worth remembering

$$E_{\theta}[s(\theta; X)] = 0$$

Under the regularity conditions listed later, the score has mean zero at the true parameter value. Its variance is therefore the natural measure of how much the data move the likelihood, and that variance is the Fisher information.

Setting the sum of the scores to zero is the likelihood equation, so the score is also the object maximum likelihood estimation solves.

RUNNING EXAMPLE

For one normal observation with known noise variance the score is linear in the observation.

$$s(\theta; x) = \frac{x - \theta}{\sigma^2}$$

A reading far from the candidate mean pushes the likelihood hard. A reading close to it says little.

Fisher information

The Fisher information is the variance of the score. It is large when the log likelihood is sharply peaked, which is exactly when the data separate nearby parameter values well.

$$I_1(\theta) = E_{\theta}\left[\left(\frac{\partial}{\partial\theta}\ln f(X; \theta)\right)^2\right] = -E_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X; \theta)\right]$$

the two forms agree under the regularity conditions, and the second is usually easier to compute

- Information is a property of the model and the parameter value, with no estimator involved.
- For independent and identically distributed data the information adds, so n observations carry n times the information of one.
- A flat log likelihood carries little information, and the bound two slides on becomes correspondingly weak.
- Information depends on how the parameter is written. Reparameterizing rescales it by the square of the derivative of the change of variable.

ADDITIVITY

With independent and identically distributed observations the total information is the single observation information times the sample size.

$$I_n(\theta) = n I_1(\theta)$$

This is why collecting more data tightens the bound in direct proportion to n .

Fisher information for the running example

Four steps, from the density to the information carried by all 25 weighings.

1

Write the log density of one observation

$$\ln f(x; \theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x - \theta)^2}{2\sigma^2}$$

2

Differentiate once in the parameter to get the score

$$s(\theta; x) = \frac{x - \theta}{\sigma^2}$$

3

Differentiate again. The curvature is constant, so no expectation is needed

$$\frac{\partial^2}{\partial \theta^2} \ln f(x; \theta) = -\frac{1}{\sigma^2}$$

4

Negate and multiply by the sample size

$$I_1(\theta) = \frac{1}{\sigma^2} = \frac{1}{225} \text{ g}^{-2} \quad I_n(\theta) = \frac{n}{\sigma^2} = \frac{25}{225} = \frac{1}{9} \text{ g}^{-2}$$

The 25 weighings together carry one ninth of an inverse gram squared of information about the mean weight.

The Cramér-Rao lower bound

No unbiased estimator in the model can have variance smaller than the reciprocal of the Fisher information. The bound is set by the model alone, before any estimator is proposed.

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

valid for every unbiased estimator of the parameter in the model

Biased estimators

The bound generalizes. The numerator picks up the derivative of the bias.

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{I_n(\theta)}$$

A biased rule can therefore sit below the unbiased bound.

Efficiency

The ratio of the bound to the achieved variance grades an unbiased estimator.

$$\text{eff}(\hat{\theta}) = \frac{1/I_n(\theta)}{\text{Var}_\theta(\hat{\theta})} \leq 1$$

An estimator that reaches one is called efficient.

Why it is useful

- It fixes a target before any estimator is built.
- An estimator that meets it is the best unbiased rule available.
- A large gap is a reason to keep looking for a better rule.

The sample mean attains the bound

Continuing the running example, with the noise standard deviation fixed at 15 g.

RUNNING EXAMPLE

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{I_n(\theta)} = \frac{\sigma^2}{n} = 9 \text{ g}^2$$

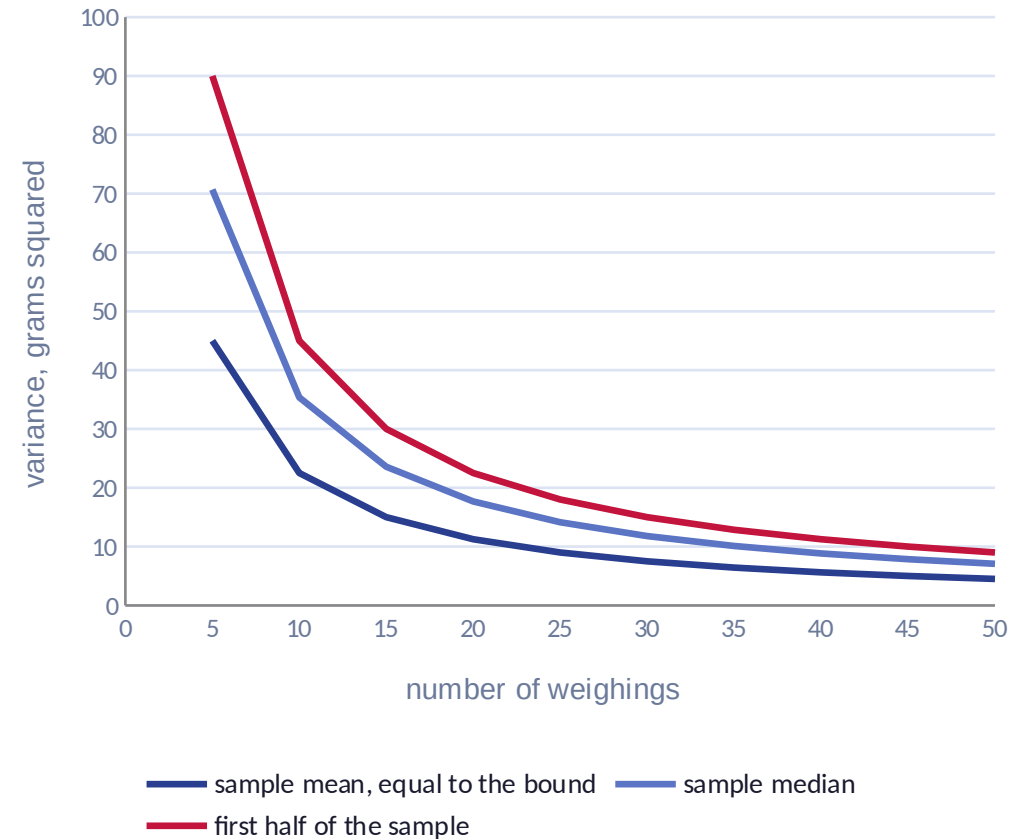
The variance of the sample mean equals the bound at every parameter value and every sample size, so its efficiency is one. It is the best unbiased estimator of the mean in this model, and no amount of cleverness will do better without accepting bias.

$$\text{Var}_\theta(\bar{X}) = 9 = \frac{1}{I_n(\theta)} \Rightarrow \text{eff}(\bar{X}) = 1$$

Two rules that do not attain it

- Averaging only the first half of the sample doubles the variance, for efficiency one half.
- The sample median is unbiased here but has variance larger by a factor of π over two, for efficiency about 0.64.

Variance of three unbiased estimators



15 Maximum likelihood and asymptotic efficiency

Maximum likelihood picks the parameter value that makes the observed data most probable. It is the default recipe for building an estimator when no obvious candidate presents itself.

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \sum_{i=1}^n \ln f(X_i; \theta)$$

solve by setting the sum of the scores to zero

Why it matters here

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta) \longrightarrow N\left(0, \frac{1}{I_1(\theta)}\right)$$

- Convergence is in distribution. The asymptotic variance is exactly the Cramér-Rao bound, so maximum likelihood is asymptotically efficient.
- It is consistent and invariant under reparameterization, which the sample mean alone does not guarantee.
- At small sample sizes it can be biased, and the asymptotic guarantee says nothing about the data actually in hand.

RUNNING EXAMPLE

For the normal model with known noise variance the likelihood equation solves in one line.

$$\sum_{i=1}^n \frac{X_i - \theta}{\sigma^2} = 0 \quad \Rightarrow \quad \hat{\theta}_{\text{ML}} = \bar{X}$$

Maximum likelihood returns the sample mean, which attains the bound at every sample size rather than only in the limit.

What the bound requires, and what it does not say

The result is standard, and so are the ways it gets misquoted.

Conditions the bound needs

- The support of the density does not depend on the parameter.
- The log density is differentiable in the parameter, and the derivative is square integrable.
- Differentiation in the parameter and integration over the data can be exchanged.
- The Fisher information is finite and strictly positive.
- The parameter is interior to the parameter space, not on a boundary.

Common misreadings

- It does not bound the mean squared error of a biased estimator. The unbiased form applies to unbiased rules only.
- A biased estimator can have smaller mean squared error than the unbiased bound, which is not a contradiction.
- Failing to attain the bound does not prove that a better estimator exists. In some models nothing attains it.
- It bounds a variance over repeated samples, not the error on the one data set in hand.
- It is silent about models that violate the regularity conditions, such as a uniform density with an unknown endpoint.

Summary

Every quantity below refers to the same model and the same estimator.

Concept	Expression	What it tells you
Loss function	$L(\theta, a) = (a - \theta)^2$	The cost charged for one guess.
Risk	$R(\theta, \hat{\theta}) = E_{\theta}[L(\theta, \hat{\theta}(X))]$	Loss averaged over every data set the model can produce.
Mean squared error	$MSE_{\theta}(\hat{\theta}) = \text{Var}_{\theta}(\hat{\theta}) + [b(\theta)]^2$	Risk under squared error loss, split into random and systematic parts.
Bayes risk	$r(\pi, \hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta$	Risk averaged over a prior, giving one number per estimator.
Minimax	$\sup_{\theta} R(\theta, \hat{\theta})$	The worst risk a rule ever incurs, minimized over rules.
Rao-Blackwell	$\tilde{\theta} = E[\hat{\theta} T(X)] \quad \text{Var}_{\theta}(\tilde{\theta}) \leq \text{Var}_{\theta}(\hat{\theta})$	Conditioning on a sufficient statistic never increases the variance.
Fisher information	$I_n(\theta) = n I_1(\theta)$	How sharply the data locate the parameter. It adds across observations.
Cramér-Rao bound	$\text{Var}_{\theta}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$	A floor on the variance of every unbiased estimator in the model.

For the running example the sample mean has risk $9 g^2$, attains the Cramér-Rao bound, and is therefore efficient.

Looking ahead to the quantum version

PREVIEW ONLY, DEVELOPED LATER IN THE COURSE

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\nu F_Q(\theta)} \quad F_Q(\theta) = \max_{\text{measurements}} I(\theta)$$

What carries over

- A parameter is encoded in a quantum state rather than in a classical density.
- Each choice of measurement produces a classical model with its own Fisher information and its own Cramér-Rao bound.
- Maximizing that information over all allowed measurements defines the quantum Fisher information.

Why it matters for this course

- The quantum bound limits what any measurement strategy can learn about a parameter, which is a statement about physics rather than about statistics.
- Entangled probes can beat the square root scaling that independent probes achieve, which is the basis of quantum metrology.
- The classical results in this deck are the tools used to prove and interpret all of it.