Primer on statistics:

MLE, Confidence Intervals, and Hypothesis Testing

Ryan Reece

ryan.reece@gmail.com
http://rreece.github.io/

Outline

1. Maximum likelihood estimators

2. Variance of MLE → Confidence intervals

3. $\Delta\chi^2$

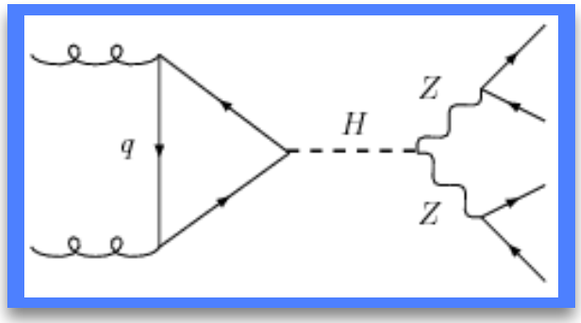
4. Efficiency example

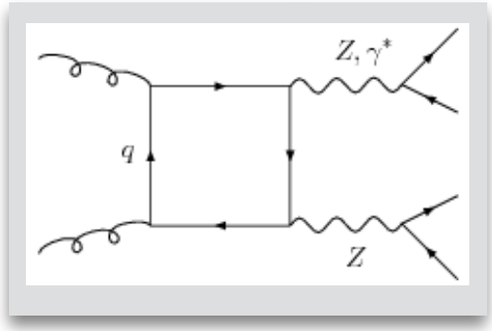
5. A/B testing

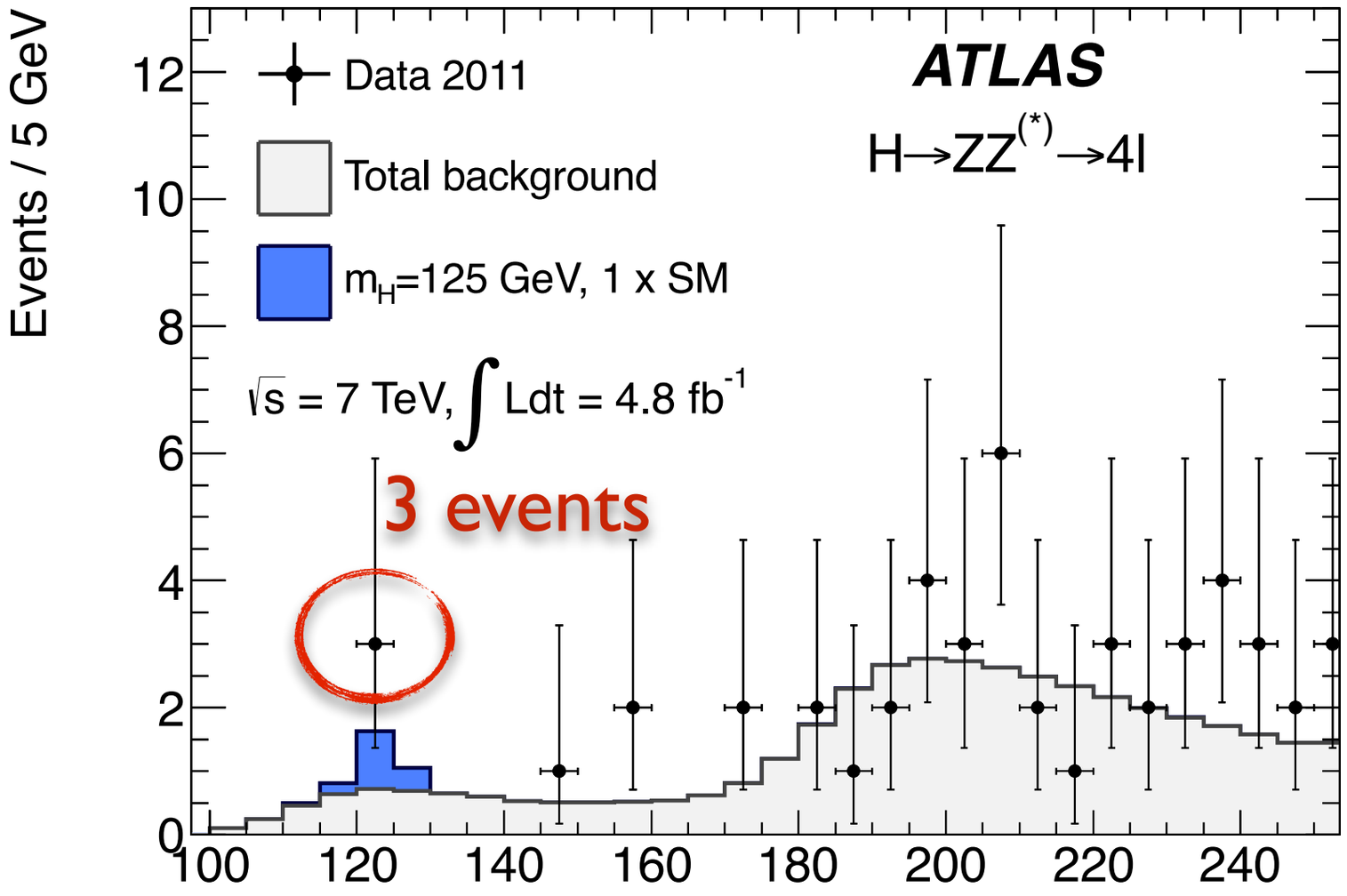
Is this significant?

Statistical questions:

- How can we calculate the best-fit estimate of some parameter?
  - ***Point estimation*** *and* ***confidence intervals***
- How can we be precise and rigorous about how confident we are that a model is wrong?
  - ***Hypothesis testing***





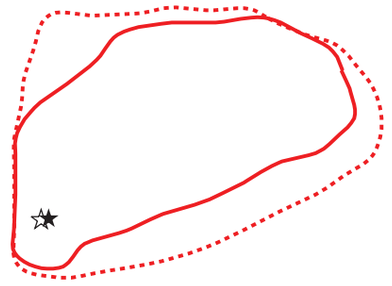


Has a local p_0 of ≈ 2%

[arxiv:1207.0319]

Confidence Intervals

- A ***frequentist confidence interval*** is constructed such that, given the model, if the experiment were repeated, each time creating an interval, 95% (or other CL) of the intervals would contain the true population parameter (*i.e.* the interval has ≈95% *coverage*).
  - They can be one-sided exclusions, *e.g.* $X > 100.2$ at 95% CL
  - Two-sided measurements, *e.g.* $X = 125.1 \pm 0.2$ at 68% CL
  - Contours in 2 or more parameters



- This ***is not the same*** as saying *"There is a 95% probability that the true parameter is in my interval"*. Any probability assigned to a parameter *strictly involves a Bayesian prior probability*.
- Bayes' theorem: P(Theory | Data) ∝ P(Data | Theory) P(Theory)

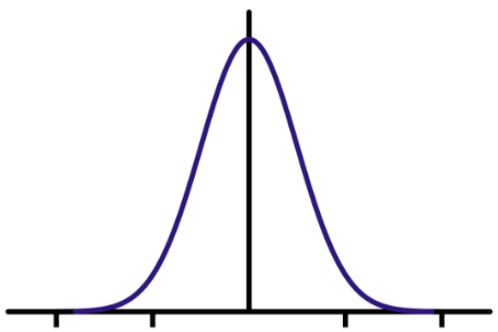
"likelihood" "prior"

Maximum likelihood method

Maximum likelihood

Consider a Gaussian distributed measurement:

$$f_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



If we repeat the measurement, the joint PDF is just a product:

$$f(\vec{x}|\mu,\sigma) = \prod_i f_1(x_i|\mu,\sigma)$$

The **likelihood function** is the same function as the PDF, only thought of as a function of the parameters, given the data. The experiment is over.

$$L(\mu,\sigma|\vec{x}) = f(\vec{x}|\mu,\sigma)$$

The **likelihood principle** states that the best estimate of the true parameters are the values which maximize the likelihood.

Maximum likelihood

It is often more convenient to consider the log likelihood, which has the same maximum.

$$\ln L = \ln \prod f_1 = \sum \ln f_1$$
$$= \sum_i \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

Maximize:

$$\frac{\partial \ln L}{\partial \mu} = \sum_i \frac{x_i - \hat{\mu}}{\sigma^2} = 0$$

$$\Rightarrow \qquad \sum_i (x_i - \hat{\mu}) = 0, \qquad \Rightarrow \qquad \hat{\mu} = \frac{1}{N} \sum_i x_i = \bar{x}$$

Which agrees with our intuition that the best estimate of the mean of a Gaussian is the sample mean.

Maximum likelihood

Note that in the case of a Gaussian PDF, maximizing likelihood is equivalent to minimizing χ^2.

$$\ln L = \sum_i \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

is maximized when

$$\chi^2 = \sum_i \frac{(x-\mu)^2}{\sigma^2}$$

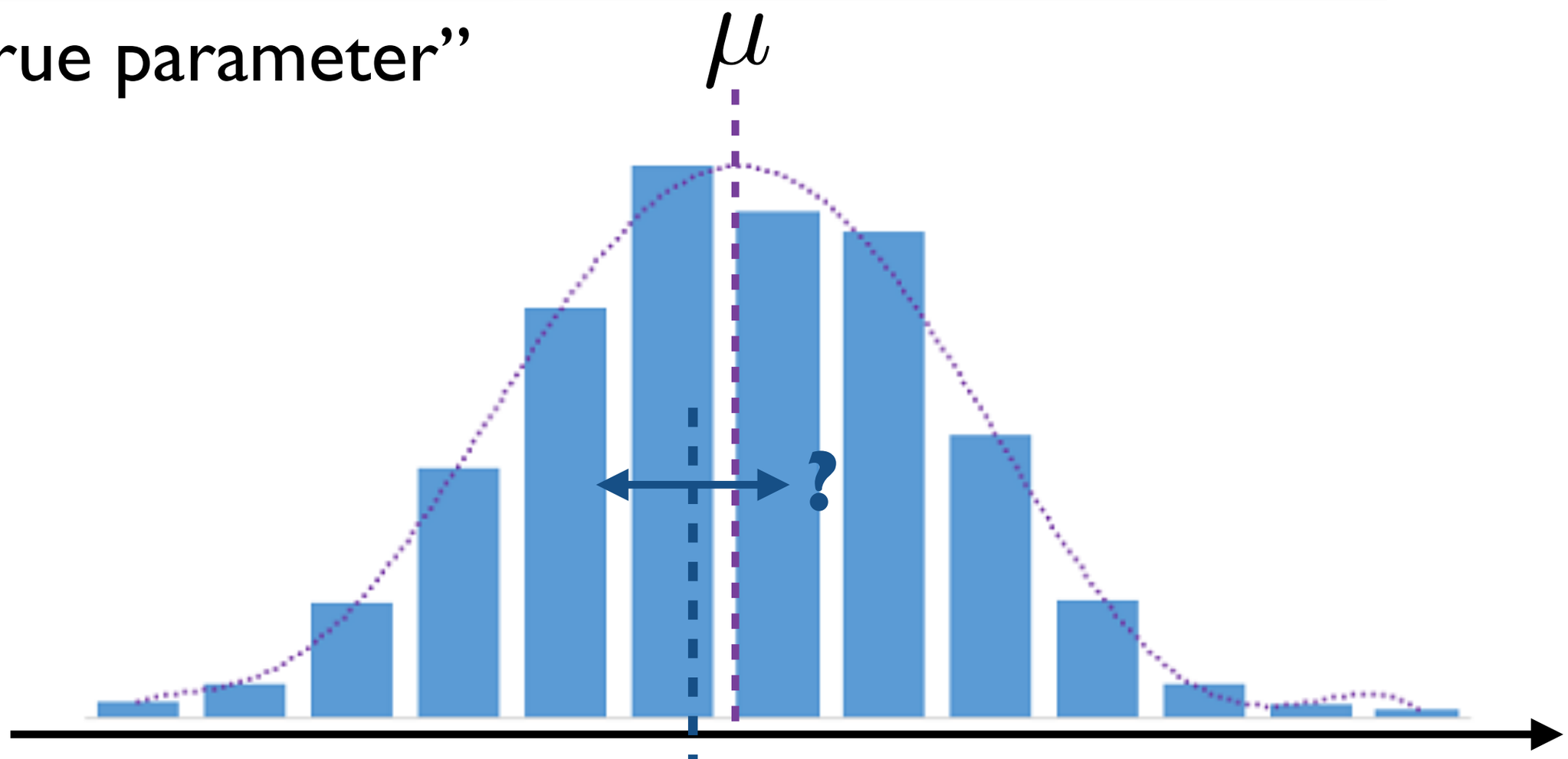
is minimized.

This was a simple example of what statisticians call **point estimation**. Now we would like to quantify our error on this estimate.

Variance of MLEs

Variance

"True parameter"



Maximum Likelihood Estimator (MLE)

$$\hat{\mu} = \frac{1}{N} \sum_i x_i = \bar{x}$$

Assuming this model, how confident are we that $\hat{\mu}$ is close to μ ? ➡ What is the variance of $\hat{\mu}$?

Variance

One would think that if the likelihood function varies rather slowly near the peak, then there is a wide range of values of the parameters that are consistent with the data, and thus the estimate should have a large error.

To see the behavior of the likelihood function near the peak, consider the Taylor expansion of a general $\ln L$ of some parameter θ, near its maximum likelihood estimate $\hat{\theta}$:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \cancelto{0}{\left.\frac{\partial \ln L}{\partial \theta}\right|_{\hat{\theta}}}(\theta - \hat{\theta}) + \frac{1}{2!}\underbrace{\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}}}_{-1/s^2}(\theta - \hat{\theta})^2 + \cdots$$

Dropping the remaining terms would imply that

$$L(\theta) = L(\hat{\theta})\ \exp\left(\frac{-(\theta - \hat{\theta})^2}{2s^2}\right)$$

Note that the $\ln L(\theta)$ is parabolic.

Variance

So if this were a good approximation, we would expect that the variance of $\hat{\theta}$ would be given by

$$V[\hat{\theta}] \equiv \sigma_{\hat{\theta}}^2 = s^2 = -\left(\left. \frac{\partial^2 \ln L}{\partial \theta^2} \right|_{\hat{\theta}} \right)^{-1}$$

It turns out that there is more truth to this than you would think, given by an important theorem in statistics, the **Cramér-Rao Inequality**:

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta} \right)^2 / E\left[- \left. \frac{\partial^2 \ln L}{\partial \theta^2} \right|_{\hat{\theta}} \right]$$

An estimator's **efficiency** is defined to measure to what extent this inequality is equivalent:

$$\varepsilon[\hat{\theta}] \equiv \frac{1/E\left[- \left. \frac{\partial^2 \ln L}{\partial \theta^2} \right|_{\hat{\theta}} \right]}{V[\hat{\theta}]}$$

Variance

It can be shown that in the large sample limit:

Maximum likelihood estimators are unbiased and 100% efficient.

Therefore, in principle, one can calculate the variance of an ML estimator with

$$V[\hat{\theta}] = -\left(E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\hat{\theta}} \right] \right)^{-1}$$

Calculating the expectation value would involve an analytic integration over the PDFs of all our possible measurements, or a Monte Carlo simulation of it. In practice, one usually uses the observed maximum likelihood estimate as the expectation.

$$\boxed{V[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\hat{\theta}} \right)^{-1}}$$

Variance

Let's go back to our simple example of a Gaussian likelihood to test this method of calculating the ML estimator's variance.

$$V[\hat{\mu}] = -\left(\left. \frac{\partial^2 \ln L}{\partial \mu^2} \right|_{\hat{\mu}} \right)^{-1}$$

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial \mu^2} &= \frac{\partial^2}{\partial \mu^2} \sum_i \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= \frac{\partial}{\partial \mu} \sum_i \frac{x_i - \hat{\mu}}{\sigma^2} = \sum_i \frac{-1}{\sigma^2} = \frac{-N}{\sigma^2} \end{aligned}$$

$$\Rightarrow \qquad V[\hat{\mu}] = \frac{\sigma^2}{N} \qquad \Rightarrow \qquad \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{N}}$$

Which many of you will recognize as the proper error on the sample mean. If you are unfamiliar with it, we can actually derive it analytically in this case.

Variance

Analytic variance of a gaussian:

$$f_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

$$V[\bar{x}] = E[\bar{x}^2] - \cancelto{\mu}{E[\bar{x}]^2}$$

$$= E\left[\left(\frac{1}{N}\sum_i x_i\right)\left(\frac{1}{N}\sum_j x_j\right)\right] - \mu^2$$

$$= \frac{1}{N^2}\,E\left[\sum_{i\neq j} x_i\,x_j + \sum_i x_i^2\right] - \mu^2$$

$$= \frac{1}{N^2}\left(\sum_{i\neq j}\cancelto{\mu^2}{E[x]^2} + \sum_i E[x^2]\right) - \mu^2$$

To find $E[x^2]$, consider

$$V[x] = \sigma^2 = E[x^2] - E[x]^2 = E[x^2] - \mu^2$$

$$\Rightarrow \qquad E[x^2] = \sigma^2 + \mu^2$$

Variance

$$\Rightarrow \qquad V[\bar{x}] = \frac{1}{N^2}\left(\sum_{i\neq j}\mu^2 + \sum_i(\sigma^2+\mu^2)\right) - \mu^2$$

$$= \frac{1}{N^2}\left((N^2-N)\mu^2 + N(\sigma^2+\mu^2)\right) - \mu^2$$

$$= \frac{\sigma^2}{N}$$

Which verifies the result we got from calculating derivatives of the likelihood function.

$$V[\hat{\theta}] = -\left(\left.\frac{\partial^2 \ln L}{\partial\theta^2}\right|_{\hat{\theta}}\right)^{-1}$$

In practice, one usually doesn't calculate this analytically, but instead:

- calculates the derivatives numerically, or
- uses the $\Delta \ln L$ or $\Delta\chi^2$ method, described now

Variance by $\Delta\chi^2$

Back to our Taylor expansion of $\ln L$:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \frac{1}{2!} \underbrace{\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}}}_{-1/\sigma^2_{\hat{\theta}}} (\theta - \hat{\theta})^2 + \cdots$$

Let $\Delta \ln L(\theta) \equiv \ln L(\theta) - \ln L(\hat{\theta})$

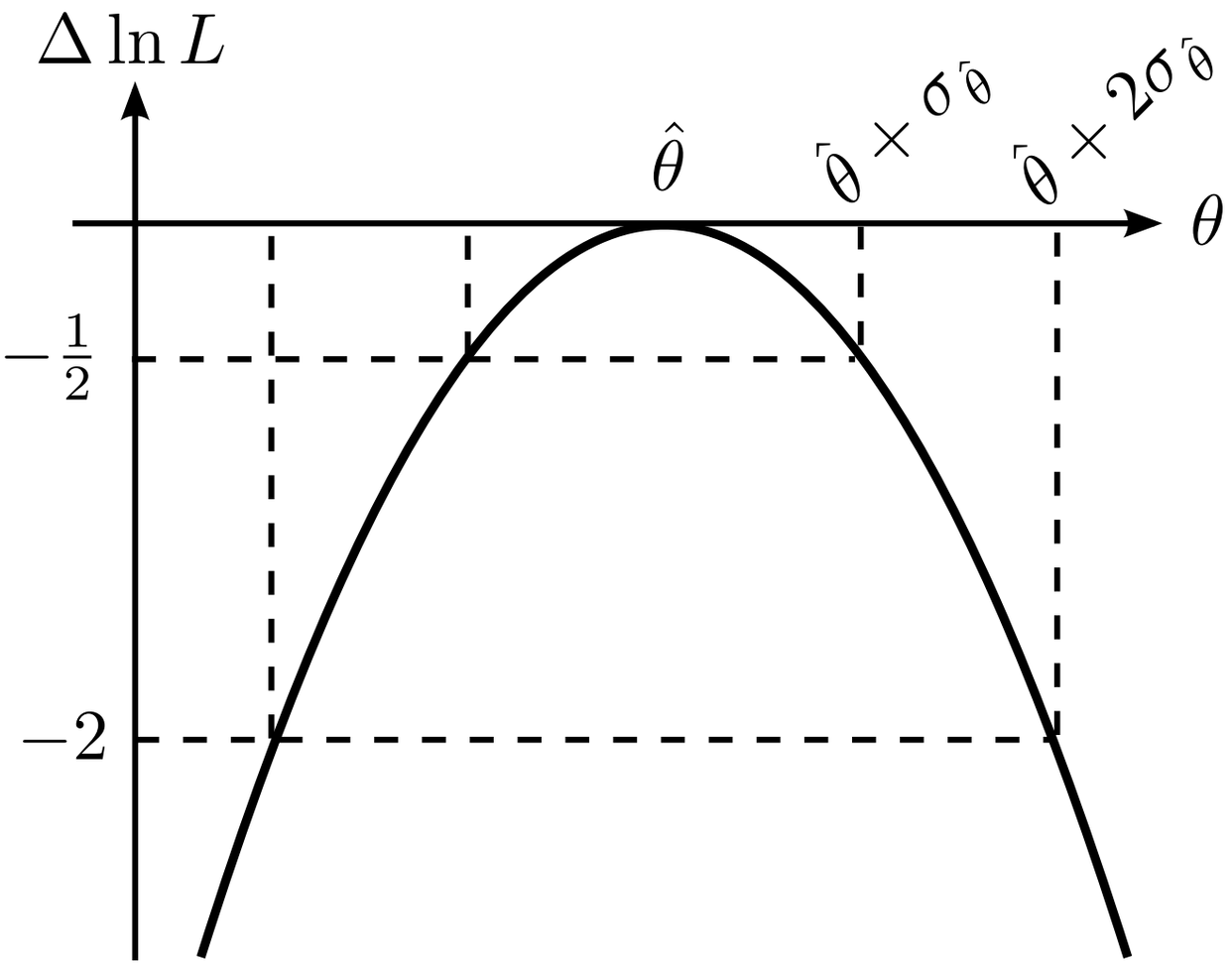
$$\Delta \ln L(\theta) \simeq -\frac{(\theta - \hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}$$

$$\theta \to \hat{\theta} \pm n\sigma_{\hat{\theta}}$$

$$\Delta \ln L(\hat{\theta} \pm n\sigma_{\hat{\theta}}) = -\frac{(\pm n\sigma_{\hat{\theta}})^2}{2\sigma^2_{\hat{\theta}}}$$

$$\boxed{\Delta \ln L(\hat{\theta} \pm n\sigma_{\hat{\theta}}) = -\frac{n^2}{2}}$$

Variance by $\Delta\chi^2$



$$\Delta \ln L(\hat{\theta} \pm n\sigma_{\hat{\theta}}) = -\frac{n^2}{2}$$

This is the most common definition of the 68% and 95% **confidence intervals**:

68%/ 1 $\sigma_{\hat{\theta}}$: $\Delta \ln L = -\frac{1}{2}$

95%/ 2 $\sigma_{\hat{\theta}}$: $\Delta \ln L = -2$

Variance by $\Delta\chi^2$

Recall that in the case that the PDF is Gaussian, the $\ln L$ is just the χ^2 statistic.

$$\ln L = -\frac{\chi^2}{2}, \qquad \chi^2 = \sum \frac{(x-\theta)^2}{\sigma^2}$$

$$\Delta \ln L(\hat{\theta} \pm n\sigma_{\hat{\theta}}) = \ln L(\hat{\theta} \pm n\sigma_{\hat{\theta}}) - \ln L_{\max} = -\frac{n^2}{2}$$

$$\Rightarrow \qquad -\frac{1}{2}\left(\chi^2(\hat{\theta} \pm n\sigma_{\hat{\theta}}) - \chi^2_{\min}\right) = -\frac{n^2}{2}$$

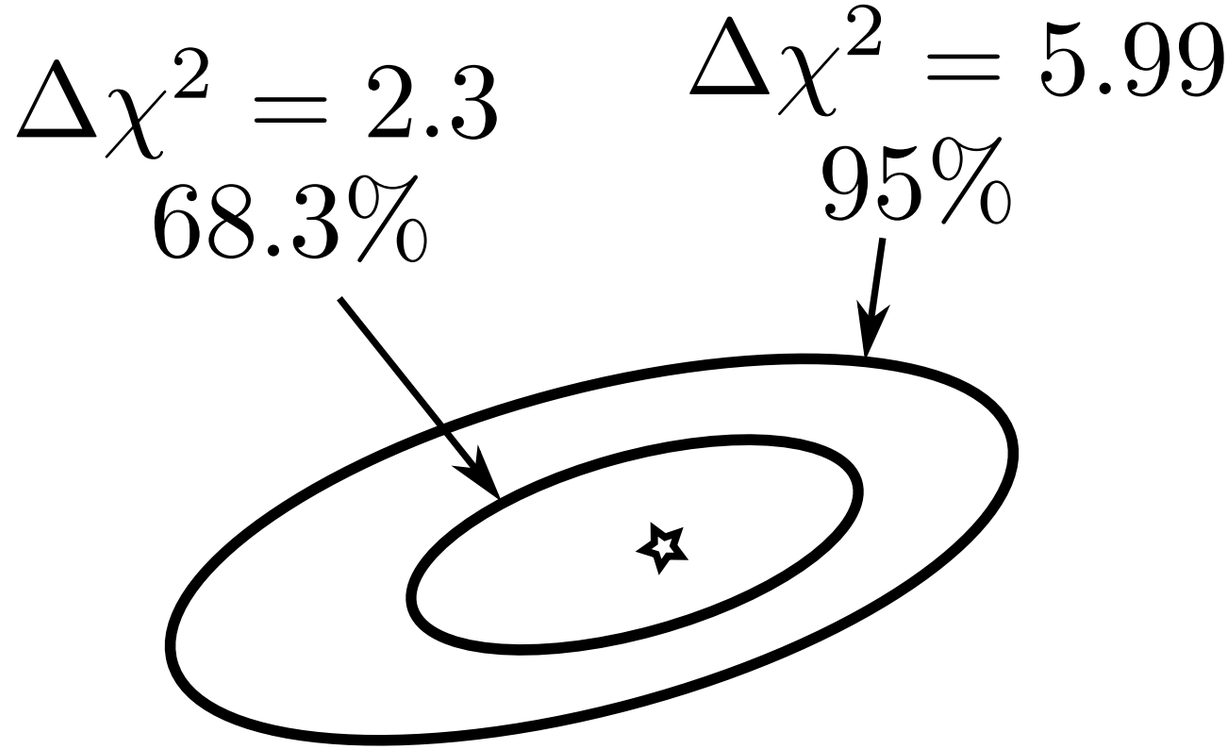
$$\boxed{\Delta\chi^2(\hat{\theta} \pm n\sigma_{\hat{\theta}}) = n^2}$$

68%/ 1 $\sigma_{\hat{\theta}}$: $\Delta\chi^2 = 1$
95%/ 2 $\sigma_{\hat{\theta}}$: $\Delta\chi^2 = 4$ (3.84 for 2σ)

Variance by $\Delta\chi^2$

Multi-dimensional case

	$Q_c = \Delta\chi^2$		
c	n=1	n=2	n=3
0.683	1.00	2.30	3.53
0.95	3.84	5.99	7.82
0.99	6.63	9.21	11.3



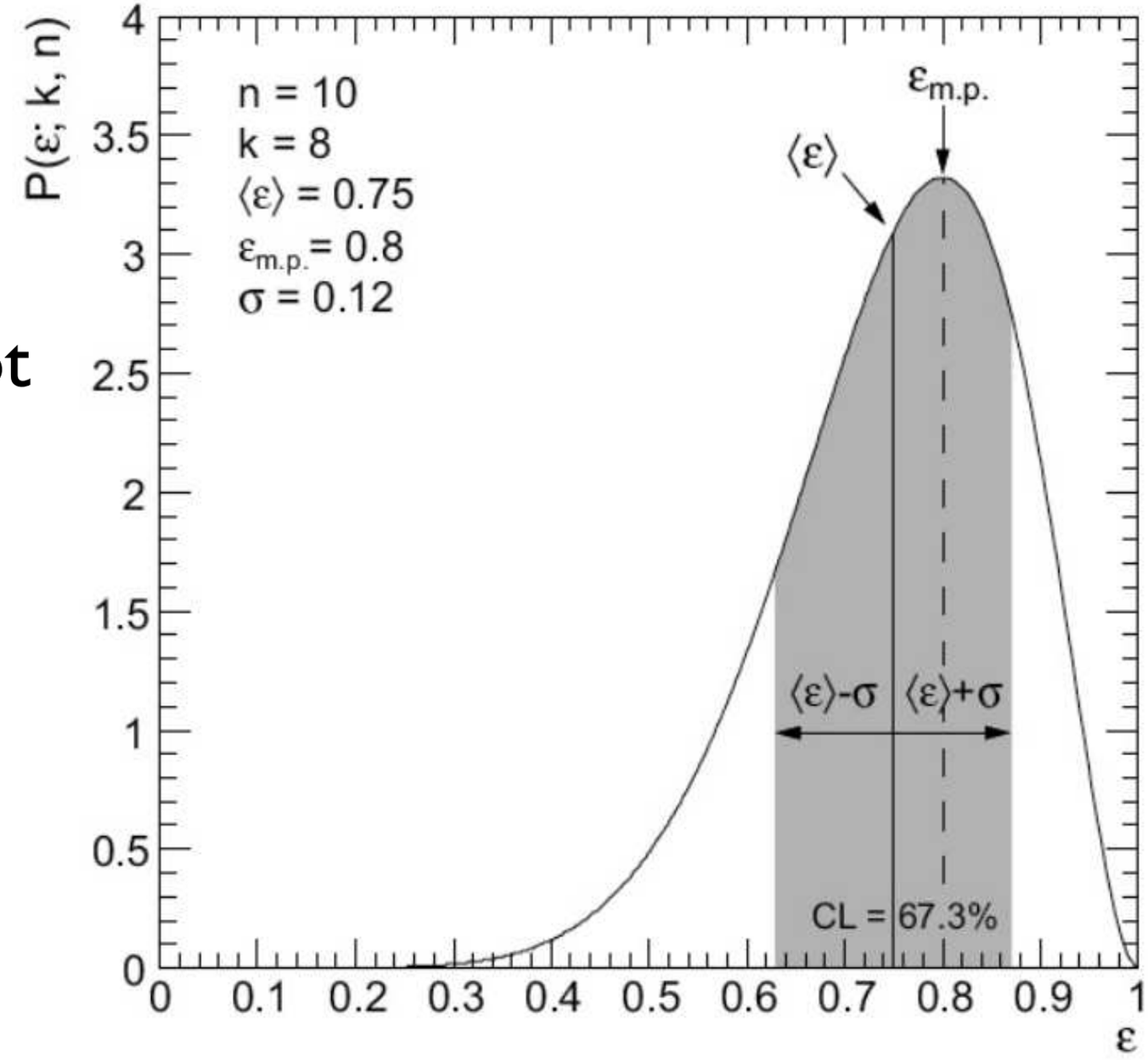
Example: measuring an efficiency & A/B testing

Efficiency

Out of *n* trials I measure *k* "conversions"

Estimate the conversion rate its precision/confidence.

Without a precision, we cannot know if observed changes are significant.



Efficiency

Binomial Distribution

MLE: $\hat{\varepsilon} = \dfrac{n}{k}$

$$
\begin{aligned}
f(k;n,p) &= \binom{n}{k} p^k \, (1-p)^{n-k} \\
{\sigma_k}^2 &= n\,p\,(1-p) \\
{\sigma_\varepsilon}^2 &= \left(\frac{\sigma_k}{n}\right)^2 = \frac{p\,(1-p)}{n}
\end{aligned}
$$

If $p \approx 0.01 \Rightarrow {\sigma_\varepsilon}^2 = \dfrac{(0.01)\,(0.99)}{n} \approx \dfrac{0.01}{n}$

$$\boxed{\sigma_\varepsilon = \frac{1}{10\sqrt{n}}}$$

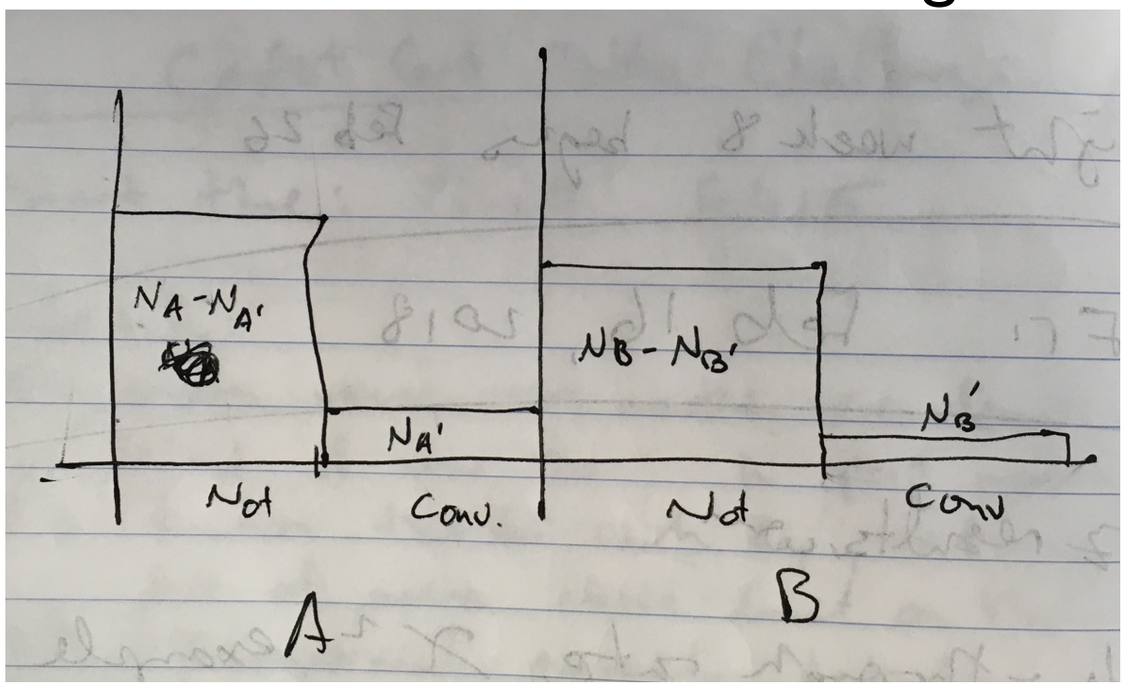
for $p = 0.01$

n	σ_ε	σ_ε / p
400	0.0050	50%
1,000	0.0032	32%
2,000	0.0022	22%
10,000	0.0010	10%
10^6	0.0001	1%

A/B testing

Example taken from here:
https://www.blitzresults.com/en/ab-tests/

Data like a 1-dim, 4-bin histogram



Assume conversion rate unchanged from A to B:

$$\varepsilon = \frac{N'_A}{N_A} \simeq \frac{N'_B}{N_B} \simeq \frac{N'_B + N'_B}{N_A + N_B}$$

Construct $\chi^2 = \sum \frac{(x-\theta)^2}{\sigma^2}$

$$\chi^2 = \frac{(N'_A - N_A\,\varepsilon)^2}{N_A\,\varepsilon} + \frac{((N_A - N'_A) - N_A\,(1-\varepsilon))^2}{N_A\,(1-\varepsilon)} + (A \to B)$$

A/B testing

	Original	Comparison variant	Sum
Visitors without Conversion	A: 960	B: 1120	2080
Visitors with Conversion	C: 40	D: 80	120
Sum	1000	1200	2200

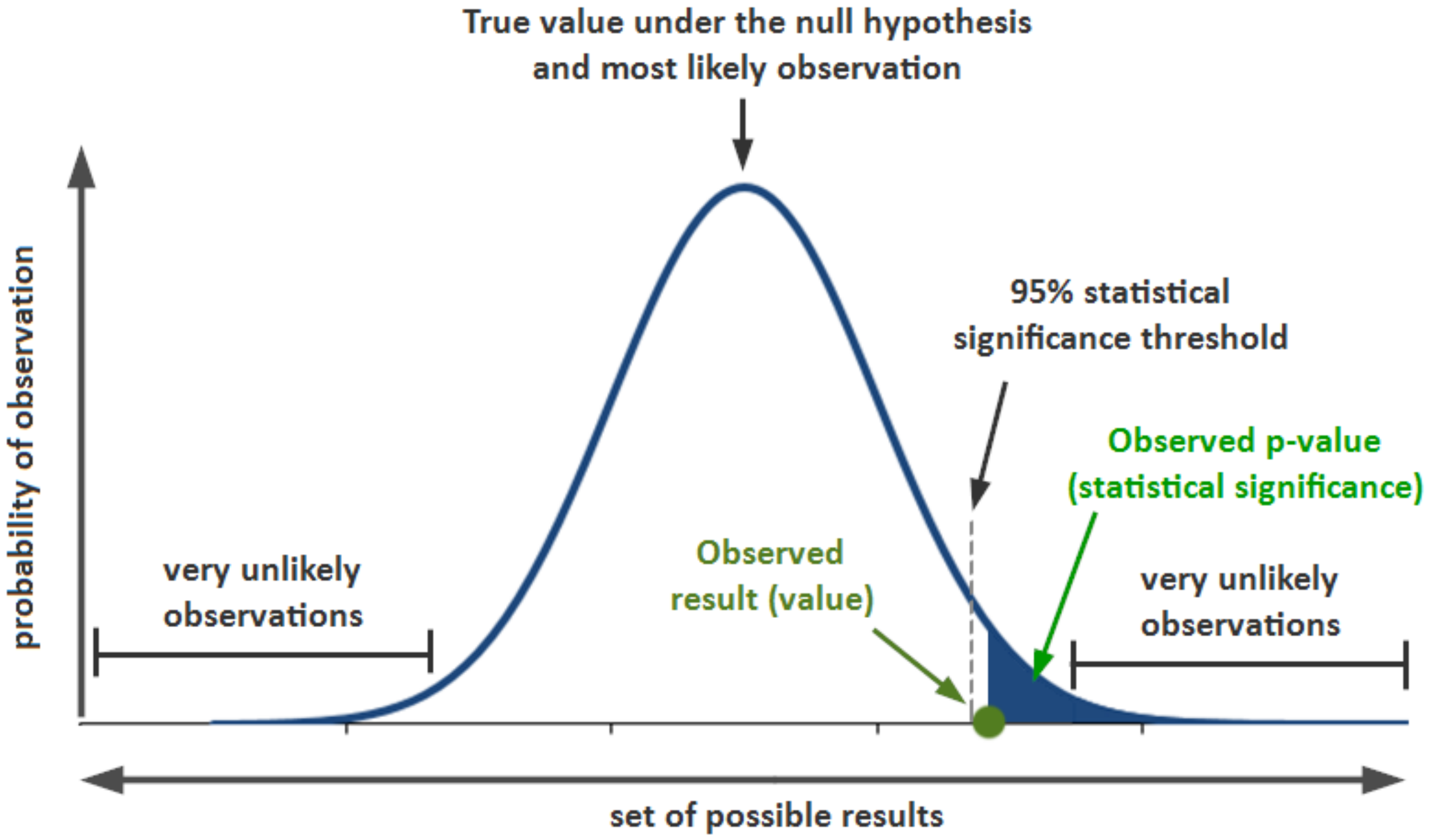
Plugin above values gives: $\Delta\chi^2$ = 7.52

Remember:

1σ (68%) : 1
2σ (95%) : 3.84
3σ (99%) : 6.63

→ >99% CL
→ significant improvement in B model

Hypothesis test



Take-aways

- MLEs are the best
- Don't just calculate MLE, find its variance!
- Quantify significance with a confidence interval
- Under many common assumptions

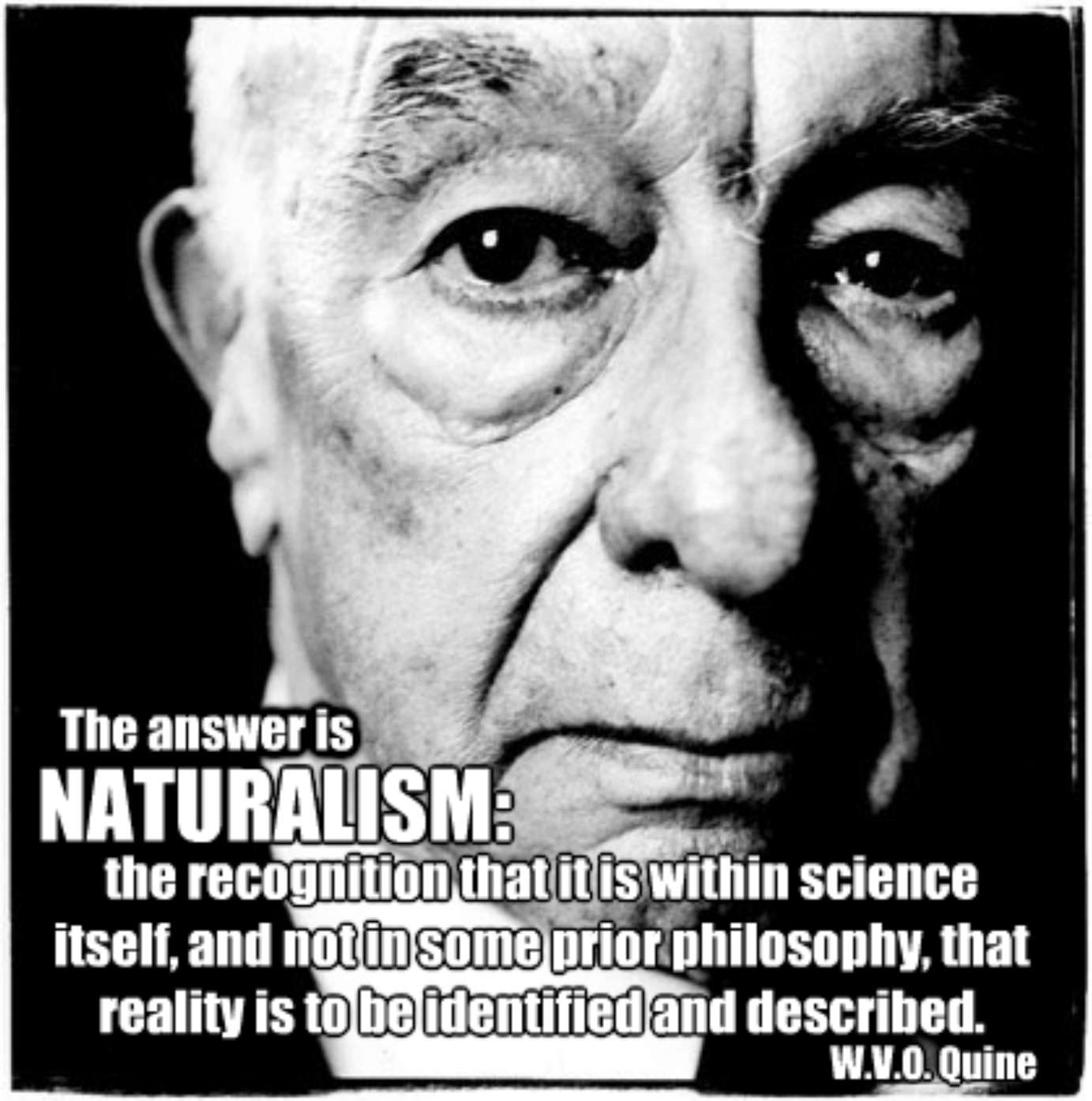
$$\ln L = -\frac{\chi^2}{2}, \qquad \chi^2 = \sum \frac{(x-\theta)^2}{\sigma^2}$$

and one can calculate $\Delta\chi^2$ to determine confidence intervals/contours.

	$\Delta\chi^2$
1σ (68%)	1
2σ (95%)	3.84
3σ (99%)	6.63

- In the simplest case of measuring an efficiency, A/B-testing amounts to a 4-term χ^2 that can be calculated by hand.
- If the $\Delta\chi^2$ is large, the change is significant!

Back up slides



Efficiency

Other examples with numbers:

Method	Numerator	Denominator	Mean (Mode)	Variance	Uncertainty σ
Poisson	1	45	0.0222	0.00050	0.02246
Binomial	1	45	0.0222	0.00048	0.02197
Bayesian	1	45	0.04255 (0.0222)	0.00085	0.02913

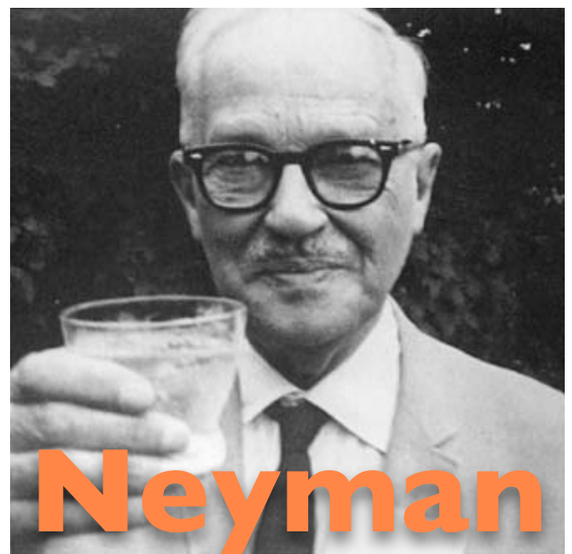
Method	Numerator	Denominator	Mean (Mode)	Variance	Uncertainty σ
Poisson	100	106	0.9433	0.01729	0.13151
Binomial	100	106	0.9433	0.00050	0.02244
Bayesian	100	106	0.9352 (0.9433)	0.00056	0.02358

Hypothesis testing

- Null hypothesis, H_0: the SM
- Alternative hypothesis, H_1: some new physics
- Type-I error: false positive rate (α)
- Type-II error: false negative rate (β)
- Power: 1-β
- Want to maximize power for a fixed false positive rate
- Particle physics has a tradition of claiming discovery at $5\sigma \Rightarrow p_0 = 2.9\times10^{-7}$ = 1 in 3.5 million, and presents exclusion with p_0 = 5%, (95% CL "coverage").
- Neyman-Pearson lemma (1933): the most powerful test for fixed α is the likelihood ratio:

$$\frac{L(x|H_0)}{L(x|H_1)} > k_\alpha$$

Table of error types		Null hypothesis (H_0) is	
		Valid/True	**Invalid/False**
Judgment of Null Hypothesis (H_0)	**Reject**	Type I error (False Positive, α)	Correct inference (True Positive, 1-β)
	Fail to reject	Correct inference (True Negative, 1-α)	Type II error (False Negative, β)
Type I = True H_0 but reject it (False Positive) **Type II = False H_0 but fail to reject it (False Negative)**			

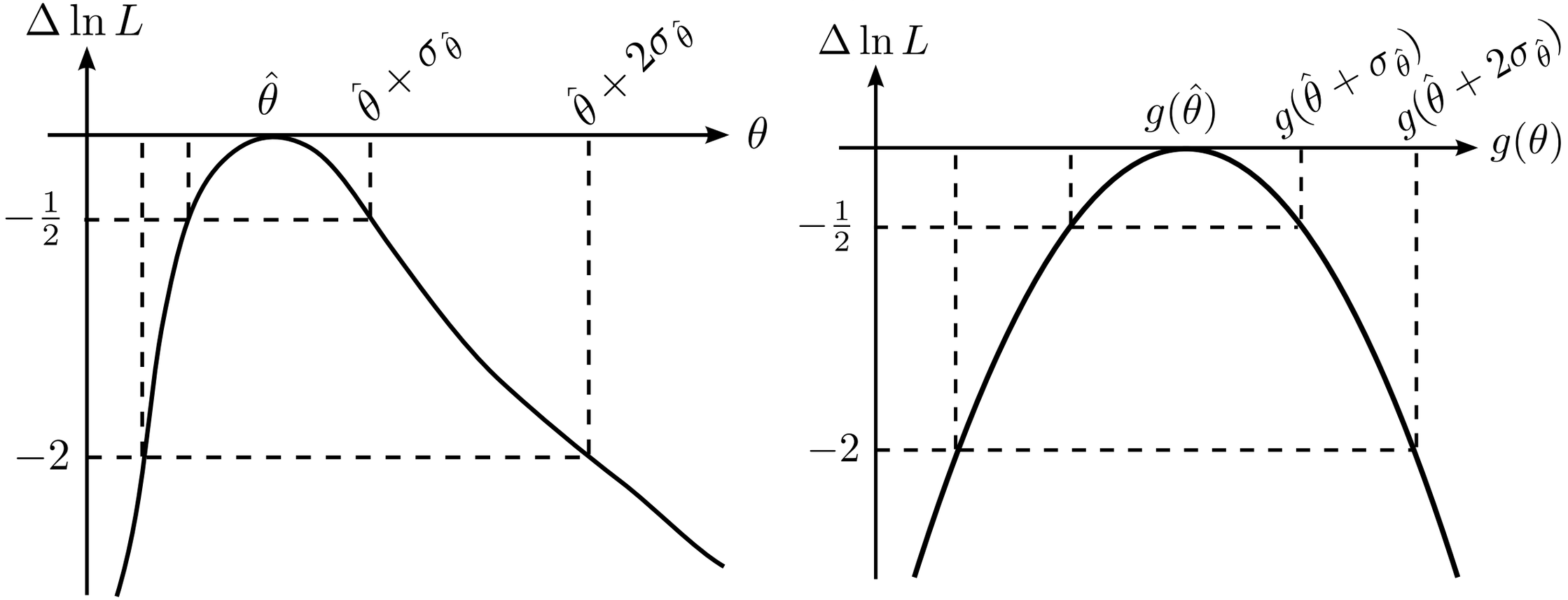


Power & Significance

> *Power: The probability of rejecting the null hypothesis when it is false. You want to design your experiment to have a power near 1.*
>
> *Significance: The probability of failing to reject the null hypothesis when it is true. You want to design your experiment so you have significance or p-values near zero.*
>
> Adapted from "The Cambridge Dictionary of Statistics", B.S. Everitt, 2nd Edition, Cambridge, 2005 printing.

Variance by $\Delta\chi^2$

What if $L(\theta)$ is not Gaussian, *i.e.* $\ln L(\theta)$ is not parabolic?



Likelihood functions have an invariance property, such that if $g(x)$ is a monotonic function, then the maximum likelihood estimate of $g(\theta)$ is $g(\hat{\theta})$. In principle, one can find a change of variables function $g(\theta)$, for which the $\ln L(g(\theta))$ is parabolic as a function of $g(\theta)$. Therefore, using the invariance of the likelihood function, one can make inferences about a parameter of a non-Gaussian likelihood function *without actually finding such a transformation* [James p. 234].

Systematics

measurement uncertainty

$$X \pm (\text{Stat} \oplus \text{Syst}_1 \oplus \text{Syst}_2 \oplus \text{Syst}_3)$$

$\propto \frac{1}{\sqrt{N}}$ — How unlucky could this be?

$\propto \frac{1}{\sqrt{N}}$ — How biased could this be?

does not scale with more data

- **statistical uncertainty:** Poisson uncertainty that scales as $1/\sqrt{N}$ (for large N).
- **class-1 systematic:** constrained in auxillary measurements in the same dataset, scales as $1/\sqrt{N}$ (for large N).
- **class-2 systematic:** an uncertainy from an independent measurement that you do not control.
- **class-3 systematic:** something not accounted for in this model (hopefully negligible).

Classification proposed by Sinervo (PhyStat2003)

from: http://philosophy-in-figures.tumblr.com/