Statistical Methods for Analysis

IceCube Bootcamp Summer 2020
Alex Pizzuto & Austin Schneider

Outline

- A quick probability review
- Statistical inference in general
  - Bayesian methods
  - Frequentist methods
- Examples!

Probability & Statistics

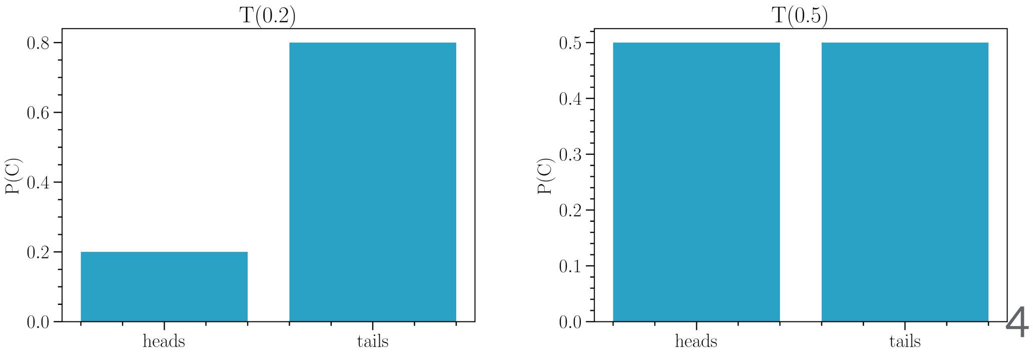
Probability

- Consider a coin.
- The probability of landing on heads or tails is 50%
- Suppose the value **C** depends on the outcome of the coin toss, then **C** is a *random variable*
- The random variable **C** is governed by a probability distribution, which we call **T** and can be written as **T(w)**
- Where **T(w)** represents the weighted coin distribution and is parameterized by **w**

$$\mathcal{P}(C = \text{heads}) = 0.5$$

$$\mathcal{P}(C = \text{tails}) = 0.5$$

$$C \sim T(w)$$



Discrete Distributions

- The T(w) distribution in the previous slide also goes by another name, the Bernoulli distribution, and is used to describe the discrete probability of a binary result (0/1 or heads/tails)

$$X \sim Ber(\beta) \implies \begin{cases} \mathcal{P}(X=1) = \beta \\ \mathcal{P}(X=0) = 1-\beta \end{cases}$$

- Bernoulli processes are rare in nature. A more common process is one governed by the Poisson distribution, which describes the probability of seeing **X** events in a given period of time.

$$\mathcal{P}(X=k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

- In this case if we examine an interval **t** with an event rate **r**, then we expect **λ=rt** events in the interval.

Discrete Distributions (cont.)

- All probability distributions are normalized

$$\sum_{k=0}^{\infty} \mathcal{P}(X = k) = 1$$

- The mean of a distribution (sometimes called average, but that's less precise), found by summing over all possible outcomes

$$\mu = \sum X \mathcal{P}(X)$$

- The act of summing over all outcomes is called calculating an expected value

$$\mu = E[X]$$

- The variance is a measure of "spread" of distributions, it is the average squared amount that the random variable *X* deviates from the mean

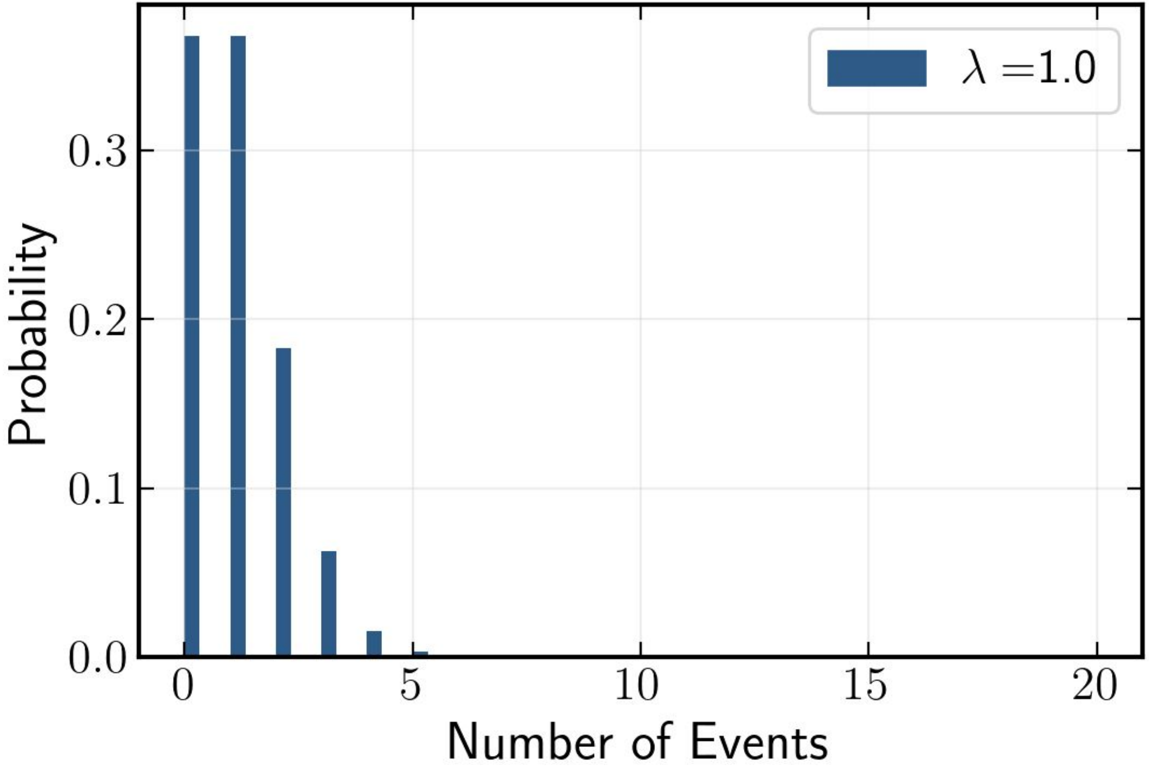
$$\sigma^2 = E[(X - \mu)^2]$$

Poisson Distribution

$$\mathcal{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\mu = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = \lambda$$

$$\sigma^2 = \lambda$$



Continuous Distributions

Discrete distributions defined over countable sets. **Continuous distributions** are defined on uncountably infinite sets, usually the set of real numbers

Our normalization condition is thus

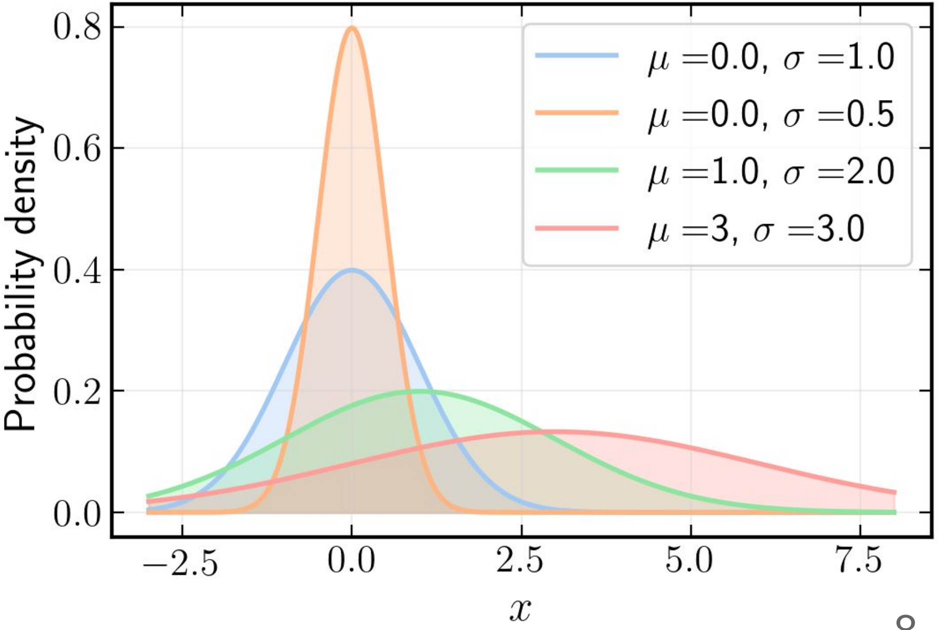
$$\int_S p(x)dx = 1$$

where *if x has units of L, then p(x) has units of 1/L.*

The most common example is the normal distribution

$$X \sim \mathcal{N}(\mu, \sigma)$$

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Joint & Conditional distributions

Suppose we have two random variables, *X* and *Y*. If these are independent, then

$$\mathcal{P}(X = x, Y = y) = \mathcal{P}(X = x)\mathcal{P}(Y = y)$$

This is the **joint probability distribution of *X* and *Y*.** If distributions are not independent, we cannot factor like this

$$p(x, y) = \frac{1}{\sqrt{6\pi}} e^{-\frac{2}{3}\left(x^2 + y^2 - xy\right)}$$

If we want to isolate the probability of one of the variables, we can **marginalize**:

$$\mathcal{P}(Y = y) = \sum_x \mathcal{P}(X = x, Y = y) \qquad p(y) = \int_{\mathcal{S}} p(x, y) dx$$

Joint and Conditional, cont'd.

If *X* and *Y* are correlated, we can decompose the joint into a product of different distributions, one separate and one **conditional**

$$\mathcal{P}(X = x, Y = y) = \mathcal{P}(X = x|Y = y)\mathcal{P}(Y = y)$$

The first term on the left is read as *probability of X = x given that Y = y*.

We can now treat the parameter of a distribution as a random variable, for example, if $X \sim \mathrm{Pois}(\lambda)$

then we can write the probability distribution as $\mathcal{P}(X = k|\lambda = L) = \dfrac{L^k e^{-L}}{k!}$

As this notation is cumbersome, we often just write

$$\mathcal{P}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Inference:
Bayesianism and Frequentism

Data, Models, and Likelihoods

- Running experiments allows us to collect data
- Proper statistical modelling is key to the interpretation of data
  - What data are you measuring?
  - What is your statistical process?
  - What distribution describes your data?
  - What is your model of the physics?
- With a model in mind, we can discuss the **probability of observing** our **data (d)** given **the model (θ)**

$$\mathcal{P}(d|\theta)$$

- Often we use a different notation, the likelihood, because we only have one set of data are more concerned with what the model parameters may be.

$$\mathcal{L}(\theta|d) = \mathcal{P}(d|\theta)$$

- Note however that the notational difference is also to avoid confusion with the probability of a parameter

$$\mathcal{L}(\theta|d) \neq \mathcal{P}(\theta|d)$$

- Generally the dependence on data is dropped for convenience

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|d)$$

Many Measurements

- Independent measurements → product of probabilities

$$\mathcal{L}(\theta) = \mathcal{P}(d_0|\theta)\mathcal{P}(d_1|\theta)$$

- Notationally we can simplify this

$$\vec{d} = (d_0, d_1, ...); \mathcal{L}(\theta) = \mathcal{P}(\vec{d}|\theta)$$

Bayesian Methods

Bayesian Inference

Model parameters are random variables

- Bayes theorem of conditional probability
- **P(θ|d)** = Posterior Distribution
- **P(d|θ)** = Likelihood
- **P(θ)** = Prior
- **P(d)** = Normalization Factor

$$\mathcal{P}(d,\theta) = \mathcal{P}(d|\theta)\mathcal{P}(\theta) = \mathcal{P}(\theta|d)\mathcal{P}(d)$$

$$\mathcal{P}(\theta|d) = \frac{\mathcal{P}(d|\theta)\mathcal{P}(\theta)}{\mathcal{P}(d)}$$

$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(\theta)\mathcal{P}(\theta)}{\mathcal{P}(d)}$$

- The “Prior” represents our prior knowledge of the parameters
- The “Likelihood” is our statistical model
- Combining these helps us to learn something about the data
- The “Posterior Distribution” is the probability of the parameters after adding knowledge of the observed data

Posterior Distribution

- The posterior distribution provide us with a picture of the parameters combining prior information and our observations

$$\mathcal{P}(\vec{\theta},\vec{\eta}|\vec{d}) = \frac{\mathcal{L}(\vec{\theta},\vec{\eta})\mathcal{P}(\vec{\theta},\vec{\eta})}{\mathcal{P}(\vec{d})}$$

- However, there are two issues
  - We don’t know the normalizing constant
  - We may not care about all the parameters
- Solution to both these problems is the same
- Integration!
- Normalization is achieved by integrating over all parameters
- Nuisance parameters **η** are integrated out
- This produces the **marginal posterior distribution**

$$\mathcal{P}(\vec{\theta}|\vec{d}) = \frac{\int d\eta\,\mathcal{L}(\vec{\theta},\vec{\eta})\mathcal{P}(\vec{\theta},\vec{\eta})}{\int d\theta\, d\eta\,\mathcal{L}(\vec{\theta},\vec{\eta})\mathcal{P}(\vec{\theta},\vec{\eta})}$$

Priors

- Say we don’t have any good external measurements for a prior
- Consider the integral on the right →
- We don’t want to be “biased”
- Maybe we choose a uniform prior (1/L) to do this

$$\int d\eta \mathcal{L}(\vec{\theta}, \eta)\mathcal{P}(\vec{\theta}, \eta)$$

$$\int d\eta \mathcal{L}(\vec{\theta}, \eta)\mathcal{P}(\vec{\theta})\frac{1}{L}$$

- Problem: what is uniform in one variable is not uniform in another
- The same prior is now not uniform

$$\phi = e^{\eta}$$

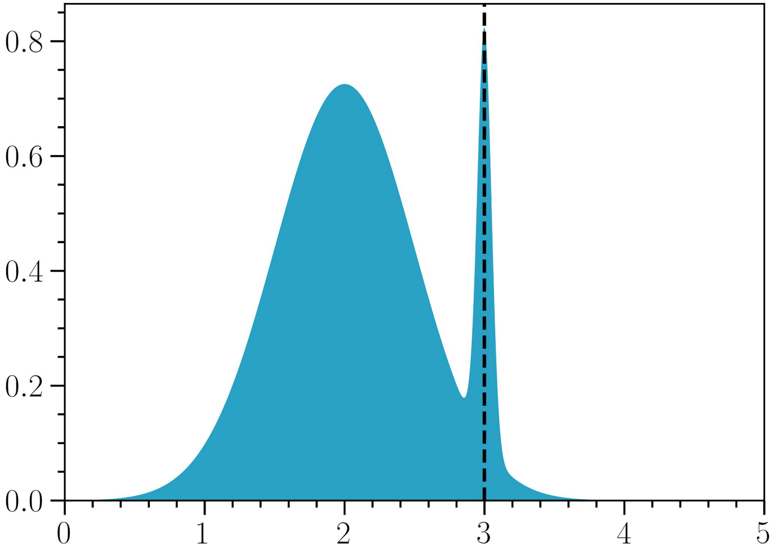
$$d\eta = \frac{d\phi}{\phi}$$

$$\int d\phi \mathcal{L}(\vec{\theta}, \phi)\mathcal{P}(\vec{\theta})\frac{1}{\phi L}$$

Parameter Estimation

- In Bayesian statistics we deal with posterior probabilities so we can ask what value of **θ** is **most probable**?
- We can select the maximum value in the marginal posterior
- This is called the **maximum *a posteriori* (MAP)** estimator

- The MAP is indicated by the dashed line
- But this is clearly not the full picture...



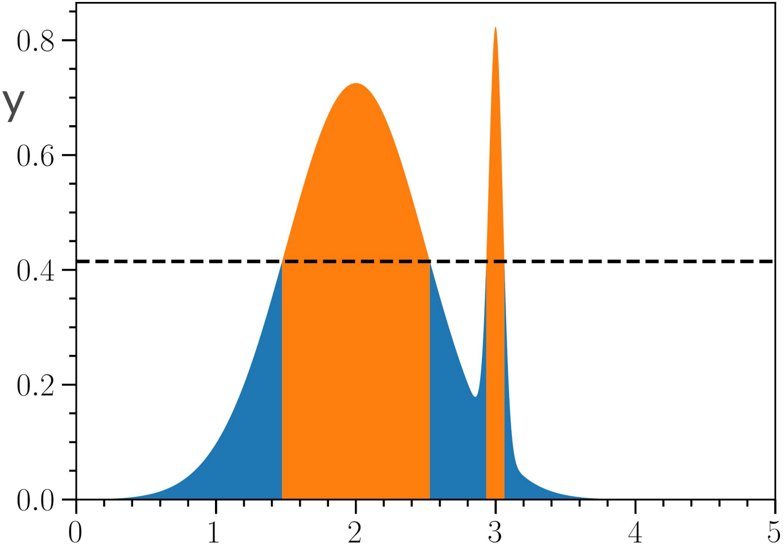
Credibility Regions

- Better to report an interval where most of the probability lies
- Want an interval C(***α***) that contains a fraction ***α*** of the probability
- Many ways to construct the interval C(***α***)

$$\alpha = \int_{\mathcal{C}(\alpha)} \mathcal{P}(\theta|d)d\theta$$

- Highest Posterior Density (HPD) is one way
- Choose only points above a threshold
- In this example ***α***=68.3%

$$\mathcal{C}_{\mathrm{HPD}}(\alpha) = \{\theta : \mathcal{P}(\theta|d) \geq f_\alpha\}$$



Model Tests

- Recall from Bayes theorem →
- Compare models by ratio of probabilities

$$\mathcal{P}(M_0|d) = \frac{\mathcal{P}(d|M_0)\mathcal{P}(M_0)}{\mathcal{P}(d)}$$

$$\mathcal{P}(M_1|d) = \frac{\mathcal{P}(d|M_1)\mathcal{P}(M_1)}{\mathcal{P}(d)}$$

$$\frac{\mathcal{P}(M_0|d)}{\mathcal{P}(M_1|d)} = \frac{\mathcal{P}(d|M_0)\mathcal{P}(M_0)}{\mathcal{P}(d|M_1)\mathcal{P}(M_1)}$$

- The ratio of P(d|M) is called the Bayes factor
- P(d|M) is the “evidence”
- The evidence is the integral of the posterior over all parameters

$$\mathcal{B}_{01} = \frac{\mathcal{P}(d|M_0)}{\mathcal{P}(d|M_1)}$$

$$\mathcal{P}(d|M_0) = \int d\theta \mathcal{P}(d|\theta, M_0)\mathcal{P}(\theta|M_0)$$

Bayesian Summary

- Model parameters are random variables
- Need priors
- Derive probability of model parameters with posterior distribution
- Parameter estimation:
  - Maximum a posteriori (MAP) estimator
- Interval construction:
  - Credibility Interval (C.I.)
  - Highest posterior density (HPD) region
- Model testing:
  - Bayes Factor
  - Evidence Integral

Computational concerns

- Computing integrals analytically can be hard
- Primary technique: Markov Chain Monte Carlo (MCMC)
- MCMCs create a list of parameter space points
- Density of parameter space points is proportional to the posterior distribution
- MAP and HPD can be computed quickly from this information
- Evidence integrals require separate but related techniques
- Sampling techniques must be chosen with care for tricky posterior landscapes
- Check that your posterior is well contained if you use improper priors!

Frequentist Methods

Frequentist Approach

Model parameters are **not** random variables

- Cannot discuss “probability of model parameters”
- Instead discuss “**probability of an outcome**” assuming these model parameters

Recall the coin toss:

$$X \sim Ber(\beta) \implies \begin{cases} \mathcal{P}(X = 1) = \beta \\ \mathcal{P}(X = 0) = 1 - \beta \end{cases}$$

If you throw a large number of coins, the fraction that have heads will be approximately ***β***

The probability of observing a fraction of heads far away from ***β*** is small

Test Statistics

- Central to frequentist statistics is the test-statistic (TS)
- The TS summarizes our observations
- Example TS for the coin toss: toss the coin n times, observe k heads
- In this case a smaller TS means closer to expectation

The χ^2 test statistic

$$\mathrm{TS}(\beta) = \sum_i^n \frac{(x_i - \beta)^2}{\beta}$$

The Likelihood Ratio TS

$$\mathcal{L}(\beta|n,k) = \frac{n!}{k!(n-k)!}\beta^k(1-\beta)^{n-k}$$

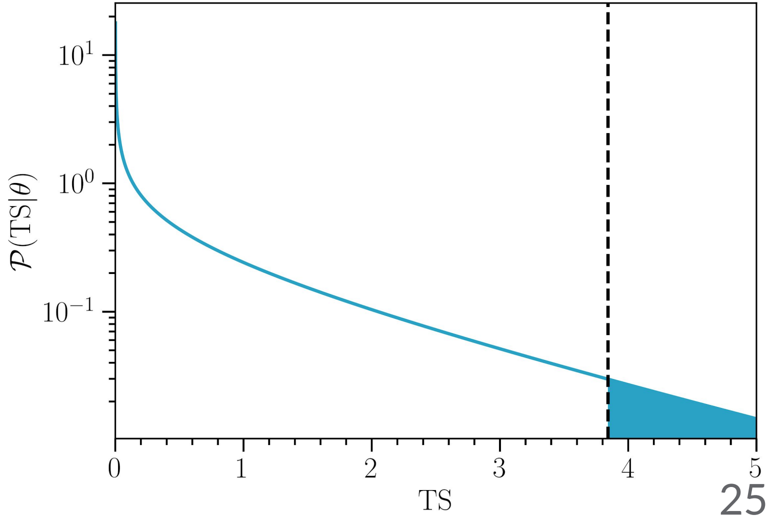
$$\mathrm{TS}(\beta) = -2\log\left(\frac{\mathcal{L}(\beta|n,k)}{\max_{\beta'}\mathcal{L}(\beta'|n,k)}\right)$$

Test Statistics

- For a particular value of the parameters there is a TS distribution
- How rare a TS value is can be answered by asking how likely a TS is to fall within a certain region
    - Usually the region greater than the observed TS

$$p_{\text{value}} = \mathcal{P}(\text{TS} > \text{TS}_{\text{obs}})$$

- Plot shows the TS for a p_{value} of 0.05
- Rare event ⇒ Small p_{value}



Parameter Estimation / Confidence Regions

- Best-fit parameters are obtained by maximizing the likelihood or equivalently minimizing the TS
- This provides an estimator for the parameters

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \mathcal{L}(\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \log \mathrm{TS}(\theta)$$

Parameter Estimation / Confidence Regions

- Intervals are constructed as observables
- For any fraction α and TS distribution, a range of TS values can be found satisfying

$$\alpha = \int_0^{\mathrm{TS}_1} \mathcal{P}(\mathrm{TS}|\theta) d\mathrm{TS}$$

$$\mathcal{P}(\mathrm{TS}_{\mathrm{obs}} > \mathrm{TS}_1) = 1 - \alpha$$

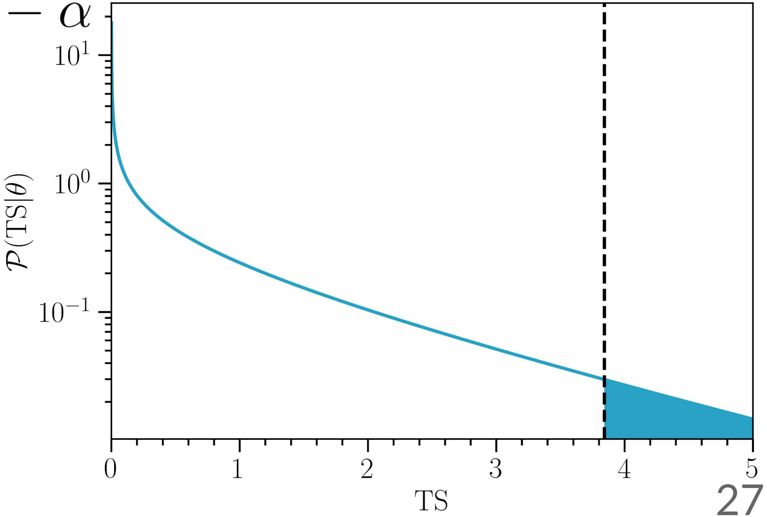
- It is possible to construct an observable with these properties

$$\gamma(\theta) = \begin{cases} 0 & \mathrm{TS}_{\mathrm{obs}}(\theta) < \mathrm{TS}_1(\theta) \\ 1 & \mathrm{TS}_{\mathrm{obs}}(\theta) > \mathrm{TS}_1(\theta) \end{cases} \left| \begin{array}{l} \mathcal{P}(\gamma(\theta) = 0) = \alpha \\ \mathcal{P}(\gamma(\theta) = 1) = 1 - \alpha \end{array} \right.$$

- As well as an interval

$$\mathcal{C} = \{\theta : \mathrm{TS}_{\mathrm{obs}}(\theta) < \mathrm{TS}_1(\theta)\}$$

- For repeated observations **C** will contain the true value of **θ** some for a fraction alpha of observations

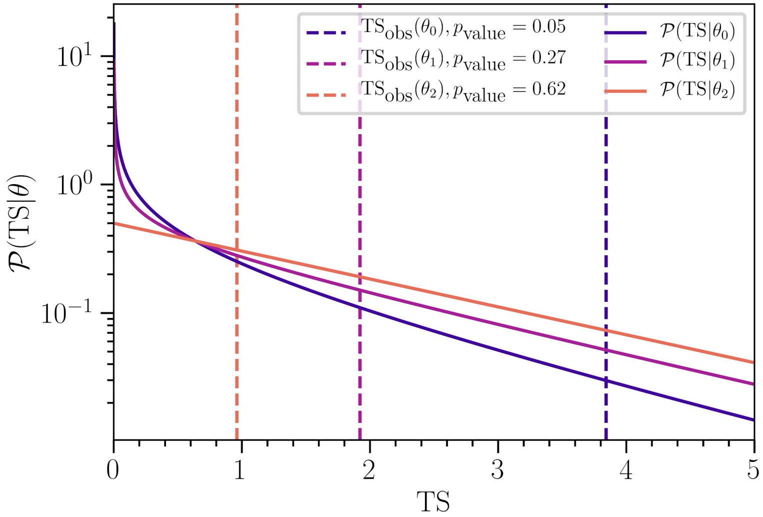


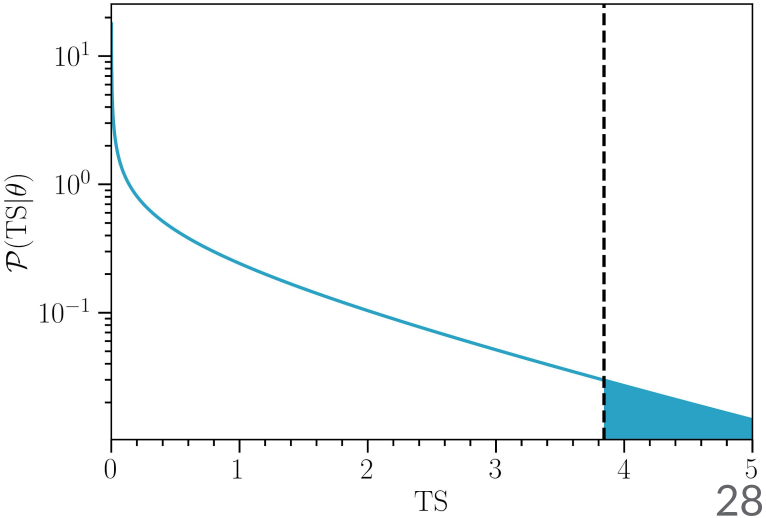
Parameter Estimation / Confidence Regions

- Note that the observed TS, TS distribution, and TS_1 depend on the parameters

$$\mathcal{C} = \{\theta : \mathrm{TS}_{\mathrm{obs}}(\theta) < \mathrm{TS}_1(\theta)\}$$

- This means it necessary to compute the TS distribution for each parameter value



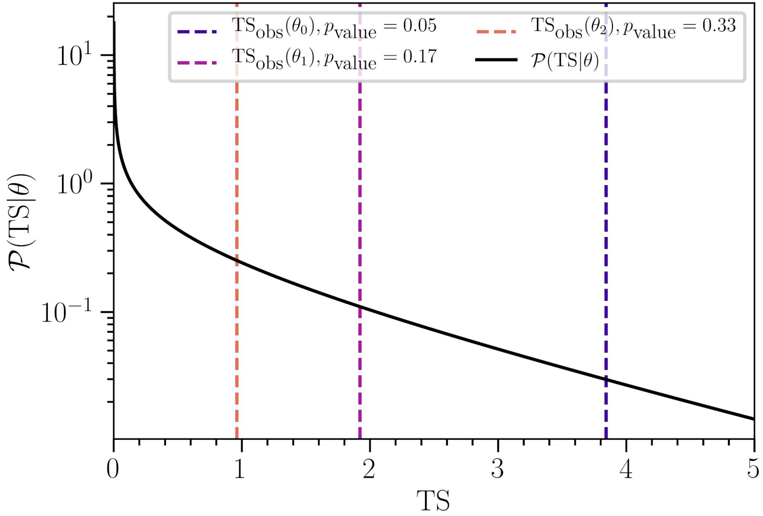


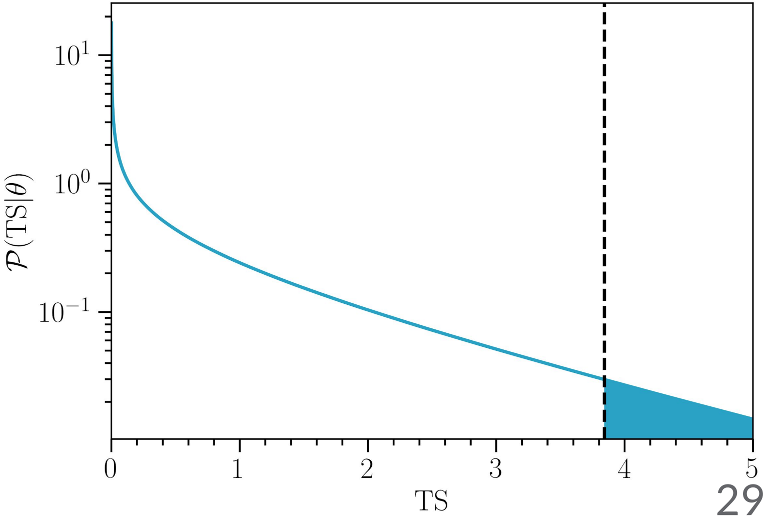
Parameter Estimation / Confidence Regions

- Note that the observed TS, TS distribution, and TS_1 depend on the parameters

$$\mathcal{C} = \{\theta : \mathrm{TS}_{\mathrm{obs}}(\theta) < \mathrm{TS}_1(\theta)\}$$

- This means it necessary to compute the TS distribution for each parameter value
- However, the TS distribution can often be approximated as χ^2 distributed if the conditions of Wilks' theorem are satisfied





Likelihood Ratios and Model Selection

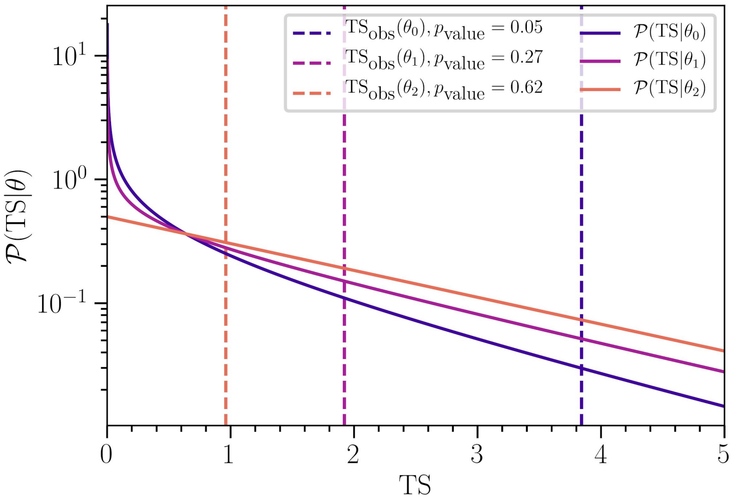
- Given two models H_0 and $H_1(\theta)$ such that H1 contains H0 as a specific point in the parameter space
- The Neyman-Pearson lemma states that the likelihood ratio test statistic has the strongest statistical power while having the lowest probability of false positives

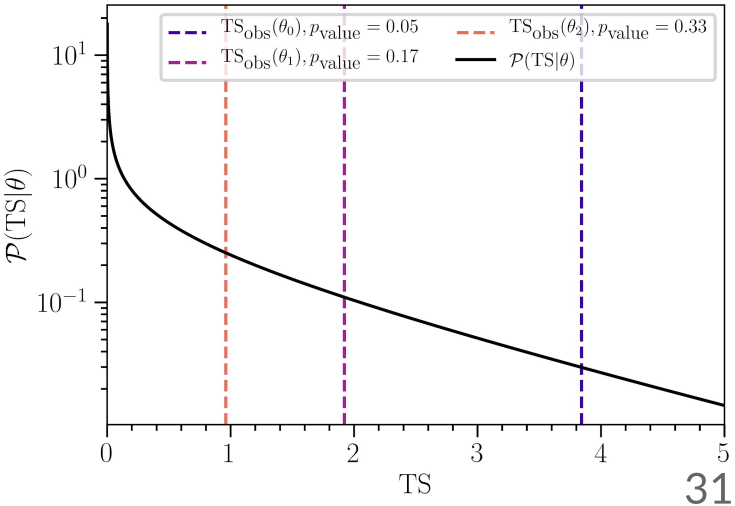
$$\mathrm{TS} = -2\log\left(\frac{\mathcal{L}(\tilde{\theta})}{\mathcal{L}(\hat{\theta})}\right)$$

- The p-value for model tests can be computed in the same manner as described previously
- By comparing the observed TS to the TS distribution of the Null hypothesis

Wilks' Theorem

- The profile likelihood test statistic is χ^2 distributed if the models satisfy the following conditions
- Asymptotic: The number of measurements tends towards infinity
- Interior: The true values of the parameters lie within the allowed bounds
- Identifiable: Each value of the parameters specifies a different model
- Nested: The null hypothesis is contained within the parameter space of the alternative hypothesis
- Correct: The model completely and correctly describes the underlying truth
- These conditions are often violated
- https://arxiv.org/abs/1911.10237 discusses when this okay, and what can be done to correct for differences or achieve a TS distribution closer to the asymptotic approximation





Frequentist Summary

- Model parameters are fixed in reality but unknown
- Data is a set of random variables
- Derive probability of observation
- Parameter estimation
    - Maximum likelihood technique
- Interval construction
    - Comparison to TS distributions
    - Probability that interval contains true parameter under repeated experiments
- Model testing
    - Likelihood ratio TS
    - Comparison to TS distribution

Computational concerns

- TS distributions can be expensive to compute for high significance results
- Number of TS distributions needed grows exponentially with the number of parameters examined
- Wilks' theorem is your saving grace

  if it applies
- Finding a global minimum can be non-trivial!
- Minimization techniques must be chosen with care for tricky TS landscapes

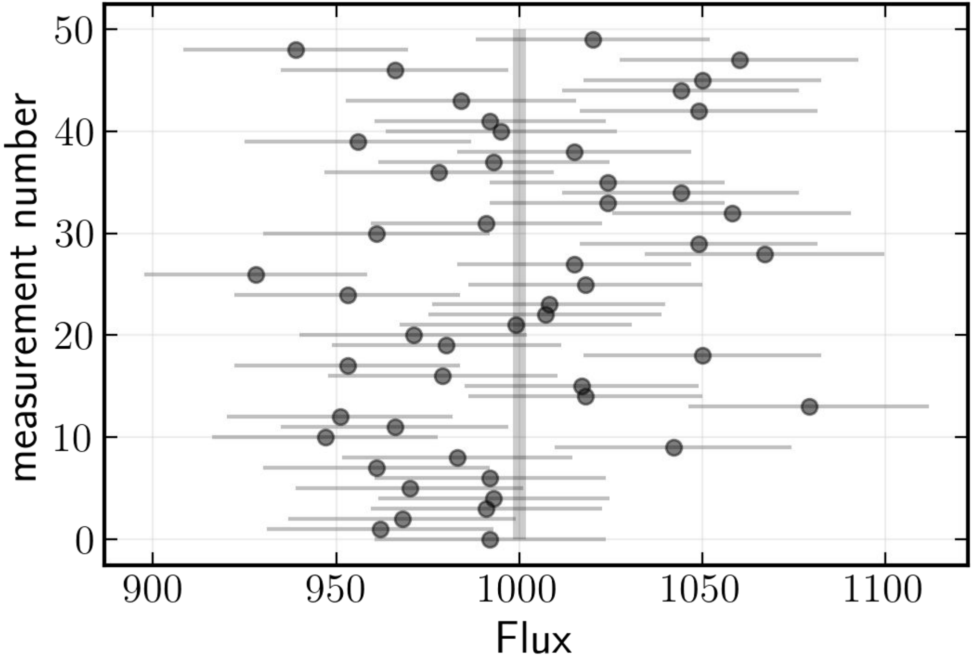
Examples

Example 1: Neutrino counts

Suppose we have a neutrino source and we want to know its flux

We take 50 independent measurements and want to make an estimate for the true flux, F_{true}

Each of our measurements, F_i, has some measurement error, e_i (assume $e_i = F_i^{1/2}$)



Example 1: Frequentist approach

We want to extremize our likelihood,

$$\mathcal{L}\left(D|F_{\text{true}}\ \right)=\prod_{i=1}^{N} P\left(D_i|F_{\text{true}}\ \right)$$

where

$$P\left(D_i|F_{\text{true}}\right)=\frac{1}{\sqrt{2\pi e_i^2}}\exp\left[\frac{-\left(F_i-F_{\text{true}}\right)^2}{2e_i^2}\right]$$

Maximizing the likelihood is the same as maximizing the log-likelihood

$$\log\mathcal{L}=-\frac{1}{2}\sum_{i=1}^{N}\left[\log\left(2\pi e_i^2\right)+\frac{\left(F_i-F_{\text{true}}\ \right)^2}{e_i^2}\right]$$

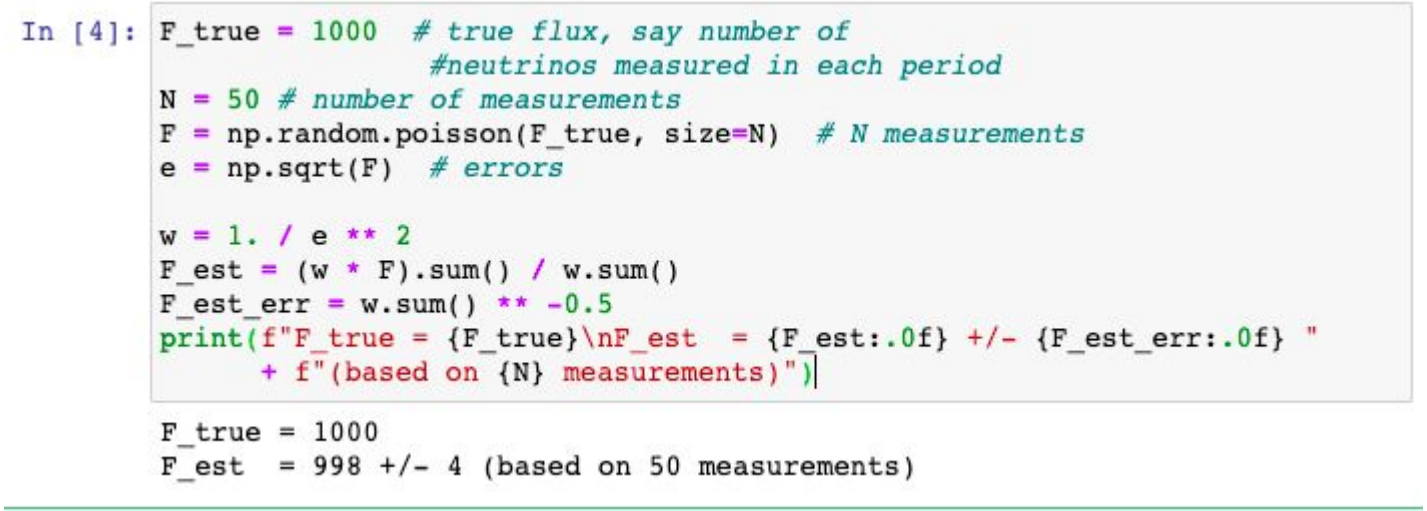
$$\frac{d\log\mathcal{L}}{dF_{true}}\Big|_{\hat{F}_{true}}=0 \quad \rightarrow \sum\frac{F_i}{e_i^2}=N\hat{F}_{true}\sum\frac{1}{e_i^2} \quad \rightarrow \hat{F}_{true}=\frac{1}{N}\frac{\sum_{i=1}^{N}F_i/e_i^2}{\sum_{i=1}^{N}1/e_i^2}$$

If the errors are equal, this just means calculate the arithmetic mean. It can be shown that the error on this estimator is

$$\sigma_{\text{est}}=\left(\sum_{i=1}^{N}1/e_i^2\right)^{-1/2}$$

Example 1: Frequentist approach

In code:



```
In [4]: F_true = 1000   # true flux, say number of
                         #neutrinos measured in each period
        N = 50 # number of measurements
        F = np.random.poisson(F_true, size=N)   # N measurements
        e = np.sqrt(F)   # errors

        w = 1. / e ** 2
        F_est = (w * F).sum() / w.sum()
        F_est_err = w.sum() ** -0.5
        print(f"F_true = {F_true}\nF_est   = {F_est:.0f} +/- {F_est_err:.0f} "
              + f"(based on {N} measurements)")
```

```
F_true = 1000
F_est  = 998 +/- 4 (based on 50 measurements)
```

Example 1: Bayesian approach

We want the maximum a posteriori probability estimator, by calculating the posterior

$$P\left(F_{\text{true}}\,|D\right)$$

To do this, we use Bayes' Theorem

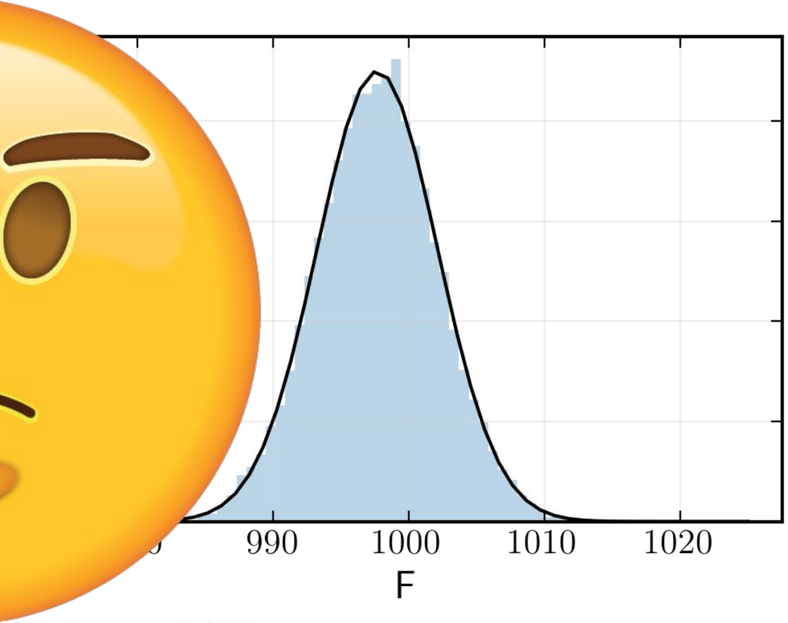
$$P\left(F_{\text{true}}\,|D\right) = \frac{P\left(D|F_{\text{true}}\,\right) P (}{P(D)}$$

Because we know nothing about the source,

uninformative prior. If we use a flat prior, ther

$$P\left(F_{\text{true}}\,|D\right) \propto \mathcal{L}\left(D|F_{\text{true}}\right.$$

But this is the exact same result we got in the freq ..



```
F_true = 1000
F_est  = 998 +/- 4 (based on 50 measurements)
```

Example 2: Bayesian Billiards

Alice and Bob enter a room with a table hidden behind a curtain

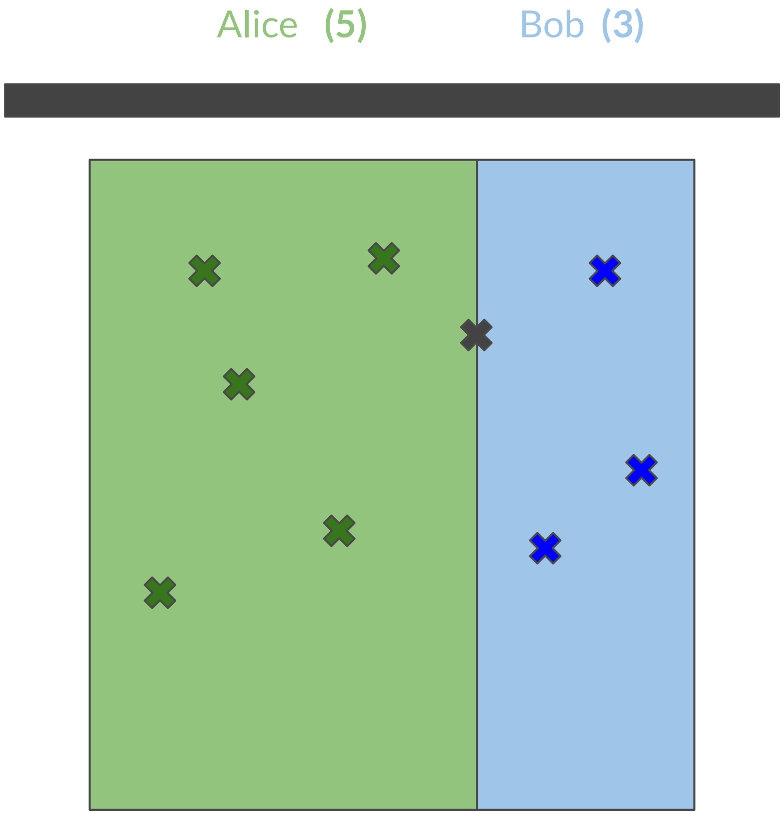
First, Carol picks a spot on the table (randomly).

Then, Carol picks other spots on the table. If they are to the left of the first spot, Alice gets a point. If they're on the right, Bob gets a point. First to 6 wins.

Assume you *do not know* where the dividing line is

Given that after 8 rolls, Alice has 5 points and Bob has 3, what is the probability that Bob will go on to win the game?

The first spot is a *nuisance parameter*: not relevant to the goal of the analysis, but necessary to take into account



Example 2: Frequentist approach

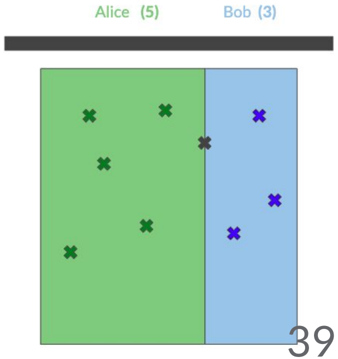
What’s the best guess at the probability that any given roll will land on Alice’s side?

$$\hat{p} = \frac{5}{8}$$

What’s the probability that Bob will win the next 3 rolls before Alice wins one (if the probability of each roll is ⅝ in favor of Alice)?

$$P(\text{Bob}) = (1 - \hat{p})^3 \approx 0.053$$

So Alice is about 18 times more likely than Bob to win



Example 2: Bayesian approach

We want to find

$$P(B|D)$$

We accomplish this by the process of *marginalization*, or integrating the joint probability, $P(B, p|D)$

$$P(B|D) = \int_0^1 P(B, p|D)dp$$

Expand the integrand using the definition of conditional probabilities

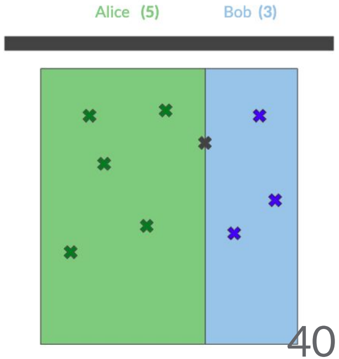
$$P(B|D) = \int P(B|p, D)P(p|D)dp$$

Use Bayes' theorem to rewrite the second term in the integrand

$$P(B|D) = \int P(B|p, D)\frac{P(D|p)P(p)}{P(D)}dp$$

And then we are left with the not-so-pretty (but very manageable)

$$P(B|D) = \frac{\int P(B|p, D)P(D|p)P(p)dp}{\int P(D|p)P(p)dp}$$



Example 2: Bayesian approach

$$P(B|D) = \frac{\int \boxed{P(B|p,D)}\,\boxed{P(D|p)}\,\boxed{P(p)}\,dp}{\int \boxed{P(D|p)}\,\boxed{P(p)}\,dp}$$

$\boxed{P(B|p,D)}$ Frequentist likelihood. Given a marker placement, *p*, and the fact Alice won 5 rolls, what is the probability that Bob will go on to six wins? $(1\text{-}p)^3$

$\boxed{P(D|p)}$ Given a probability *p*, what is the likelihood of 5 Alice wins and 3 Bob wins? $p^5(1\text{-}p)^3$

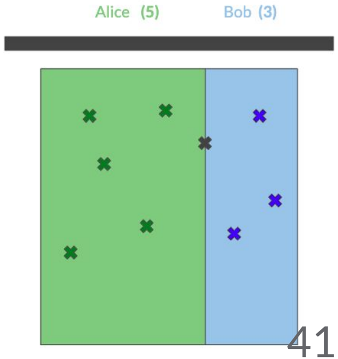
$\boxed{P(p)}$ Prior on the probability *p*. From the problem definition, *p* is evenly sampled

$$P(B|D) = \frac{\int_0^1 (1-p)^6 p^5 dp}{\int_0^1 (1-p)^3 p^5 dp}$$

After making Python do these integrals (they're Beta functions), we get

$$P(B|D) = 0.09$$

So Alice is 10 times more likely to win, not 18 times . . .



Okay, but who is *right*?

Let's check. We will play 100000 games i.e. we sample 100000 "first rolls" and recognize each game will have at most 11 more rolls

Then, figure out which games have Alice with 5 points and Bob with 3 after 8 rolls

Finally, out of these games, count how many Bob ends up winning

```
# compute the probability
mc_prob = bob_won.sum() * 1. / good_games.sum()
print("Monte Carlo Probability of Bob winning: {0:.2f}".format(mc_prob))
print("MC Odds against Bob winning: {0:.0f} to 1".format(
     (1. - mc_prob) / mc_prob))
```

```
Monte Carlo Probability of Bob winning: 0.09
MC Odds against Bob winning: 10 to 1
```

```
Number of these games Bob won: 973
```

Example 2: Discussion

The naive maximum likelihood approach (frequentist) led us astray. Why?

Frequentism is not necessarily wrong, this is more of an example of the naivety of the approach. The difficulty is that in Frequentist approaches, it is difficult to deal with nuisance parameters in non-Bayesian ways

Also, the question itself was more posed for a Bayesian answer: Frequentists might instead hope to give the answer in terms of null tests or confidence intervals, which might be classically accurate, but doesn't quite answer the question at hand.

Example 3: Confidence vs. Credibility

This example highlights the main difference between the questions Frequentists and Bayesians are answering

Frequentists: probability is a **measure of the frequency of (perhaps hypothetical) repeated events**

Bayesians: probability is a **measure of the degree of certainty about values**

As a result of this:

Frequentists: consider **model parameters to be fixed and data to be random**

Bayesians: consider **model parameters to be random and data to be fixed**

Example 3: 95% intervals for a Gaussian mean

Frequentist: Confidence interval

Unbiased estimate of the mean and standard error of the mean given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma_\mu = \sigma_x/\sqrt{N}$$

So our confidence interval is

$$CI_\mu = (\bar{x} - 2\sigma_\mu, \bar{x} + 2\sigma_\mu)$$

There is a 95% probability that when I compute CIμ from data of this sort, the true mean will fall within CIμ.

Bayesian: Credible region

It can be shown that

$$P(\mu|D) \propto \exp\left[\frac{-(\mu - \bar{x})^2}{2\sigma_\mu^2}\right]$$

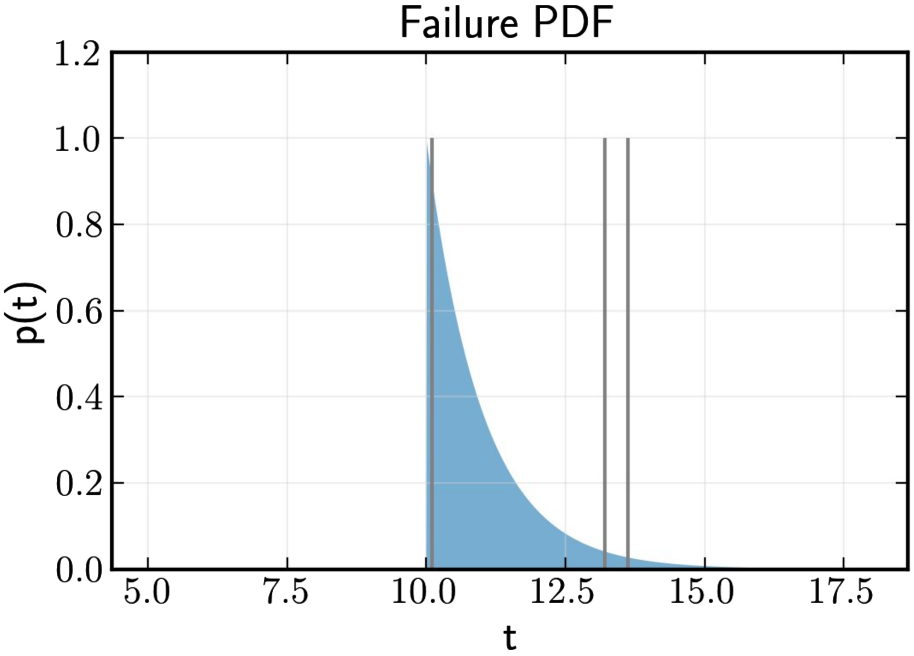
So our posterior is the same as the sampling distribution in the frequentist case. It follows that

$$CR_\mu = (\bar{x} - 2\sigma_\mu, \bar{x} + 2\sigma_\mu)$$

Given our observed data, there is a 95% probability that the true value of μ falls within CRμ

Example 4: Jaynes' Truncated Exponential

Suppose we build some DOMs for Gen-2, and cover them in protective coating that *assures full functionality for a time period*, θ (but we don't know the value of θ). After this time period, failures occur according to an exponential

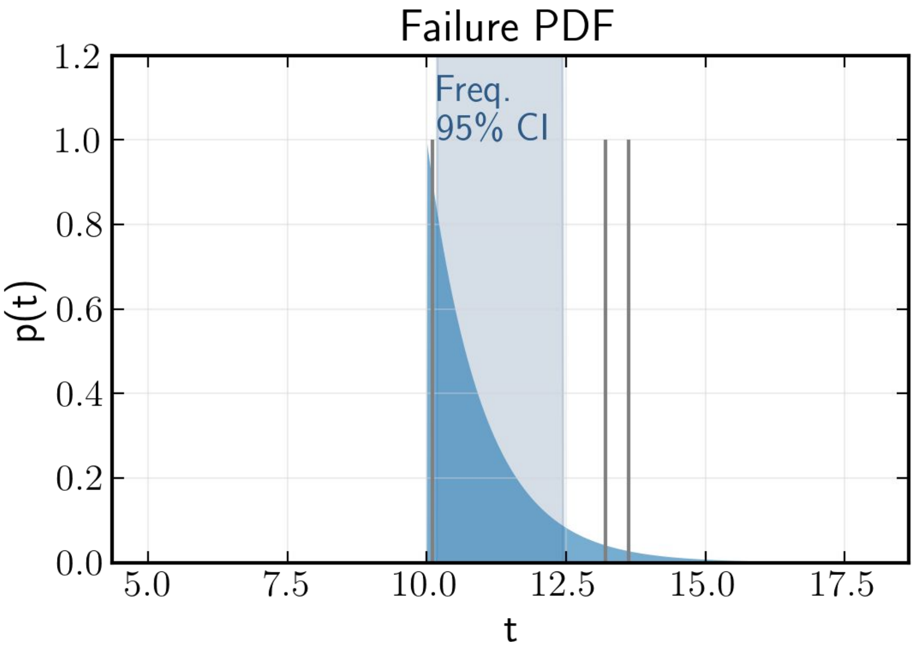


$$p(t|\theta) = \begin{Bmatrix} \exp(\theta - t) & , & t > \theta \\ 0 & , & t < \theta \end{Bmatrix}$$

Suppose we want to make a guess at what θ is, by observing some failures

Example 4: Frequentist approach

Start by noticing that we can calculate the expectation of the distribution



$$E(x) = \int_0^\infty x p(x) dx$$
$$= \theta + 1$$

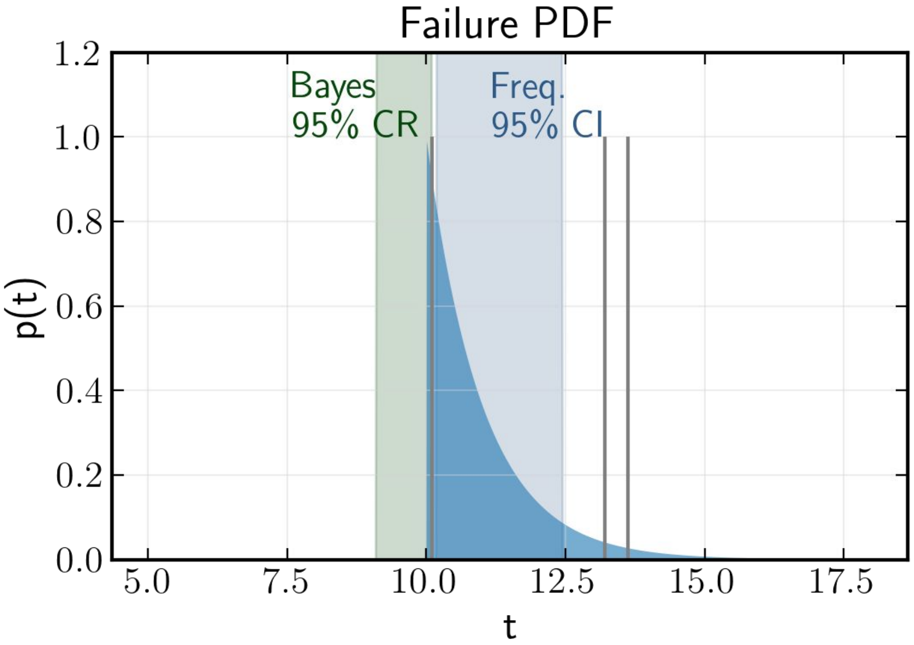
And then standard error of the mean as before

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i - 1$$

And 95% containment means +- 2 standard errors

$$CI = \left(\hat{\theta} - 2N^{-1/2}, \hat{\theta} + 2N^{-1/2}\right)$$

Example 4: Bayesian approach



For the Bayesian solution, start with Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{P(D)}$$

Plugging in our likelihood gets us

$$p(\theta|D) \propto \left\{ \begin{array}{ll} N\exp[N(\theta - \min(D))] \, , & \theta < \min(D) \\ 0 \, , & \theta > \min(D) \end{array} \right\}$$

Then, integrate to our desired level (f=0.95)

$$\int_{\theta_1}^{\min(D)} N\exp\left[N\left(\theta - \min(D)\right)\right] d\theta = f$$

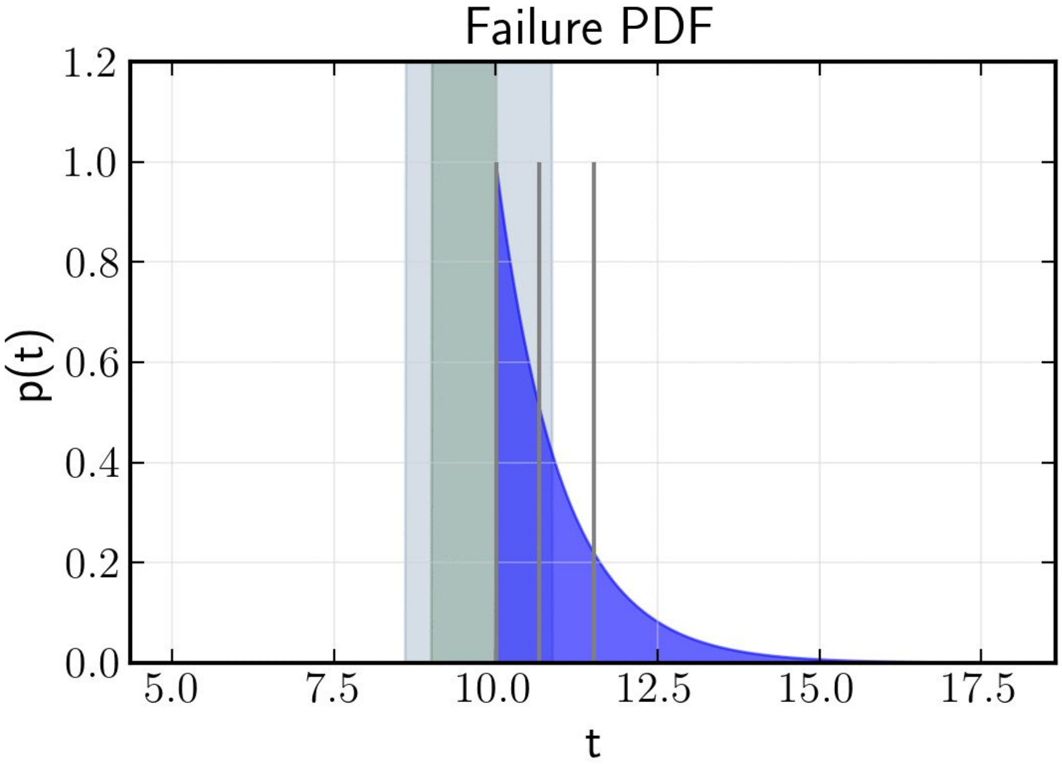
Solving this gives us our lower bound

$$\theta_1 = \min(D) + \frac{\log(1-f)}{N}$$

And our credible region is then $CR = (\theta_1, \min(D))$

Example 4: Comparisons

I’m not saying Frequentists are *wrong*, I just picked an example where they disagreed. The limits do have the proper coverage (ie in 95% of cases, the CI constructed does contain the true value)



Summary

I am **not** saying that one interpretation is better than the other. They answer different questions.

- Frequentists make statements about *ensembles of constructed intervals* (where your constructed interval is just one sample from this ensemble)
- Bayesians make *probabilistic statements about fixed intervals*.
- You **cannot** interpret a frequentist confidence interval in the same way that you would a Bayesian credible region (this is a common mistake)

When used correctly, both sets of statistical tools can be used to effectively interpret data

Both frequentist and Bayesian techniques have their own computational challenges. Be sure that you check which one is feasible for your situation.

Do you have reasonable priors available? How many parameters do you have? Is the likelihood/posterior space smooth or horrible looking? Is Wilks' theorem valid for your case? Are there other results you want to compare to? How expensive is your likelihood evaluation? What kinds of computing resources do you have access to?