STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 6: Hypothesis testing and confidence intervals (part 2)

Hypothesis testing: reminder

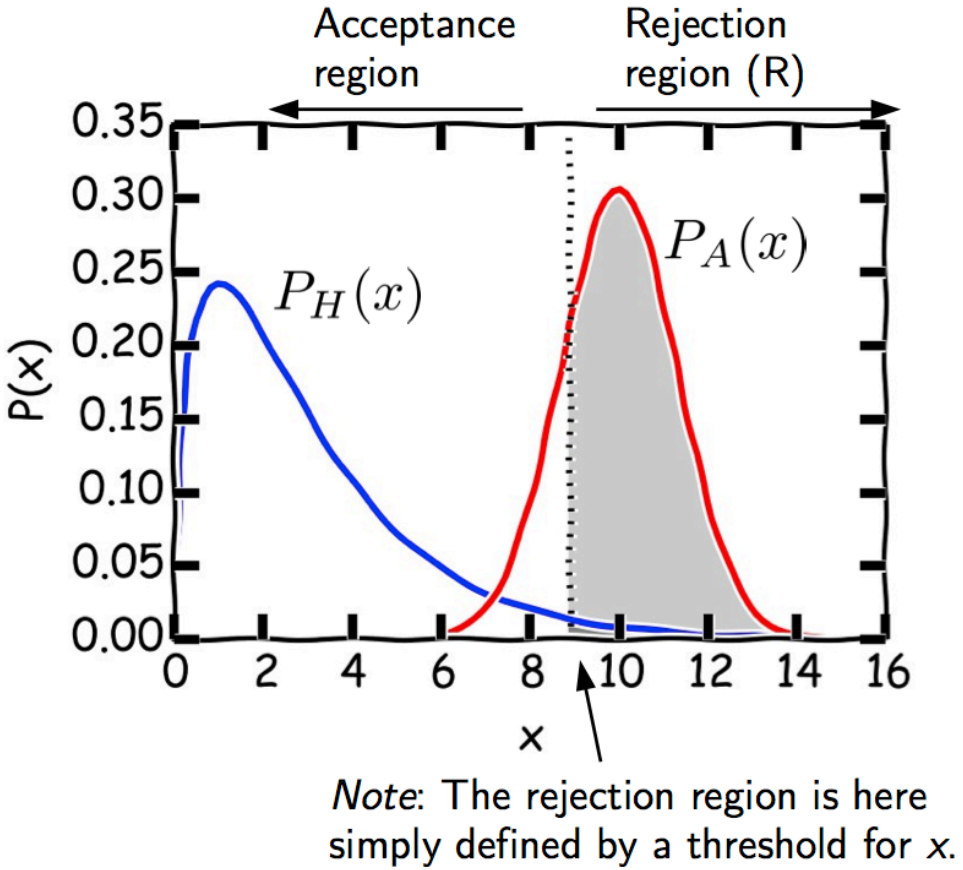
- **Statistical significance** for rejection of the null hypothesis:

$$\int_R P_H(x)dx = \alpha$$

"The observation has a *p*-value smaller than *α*"

- **Statistical power** of the test

$$\int_R P_A(x)dx = 1 - \beta$$



A good test minimises the chance for the following failure modes:

- Type I error (false positive): ***reject a true*** null hypothesis (with probability *α*)
- Type II error (false negative): ***accept a false*** null hypothesis (with probability *β*)

Maximising statistical power: likelihood ratio

- Given a desired significance level α, what is the rejection region that maximises the statistical power of a test?

Neyman-Pearson Lemma:

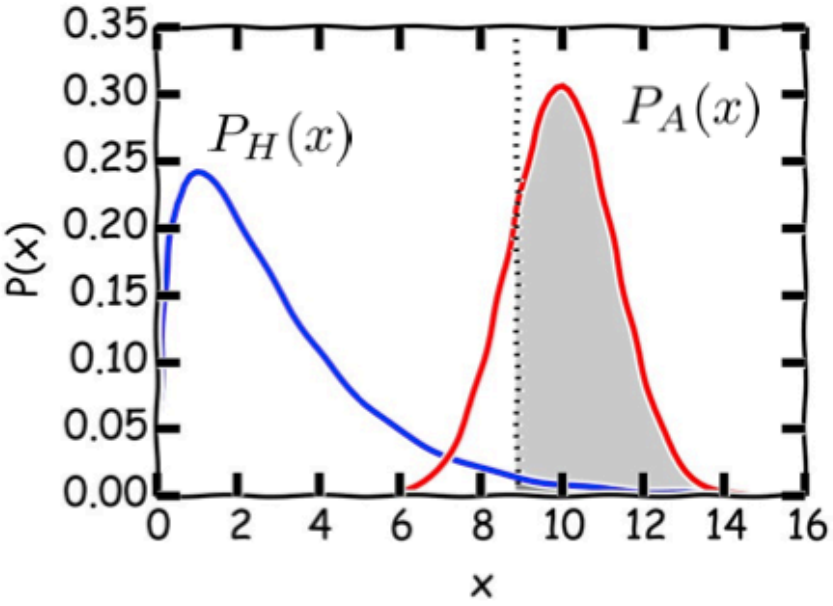
- The rejection region that maximises the statistical power is given by all *x* that have a large enough ***likelihood ratio***:

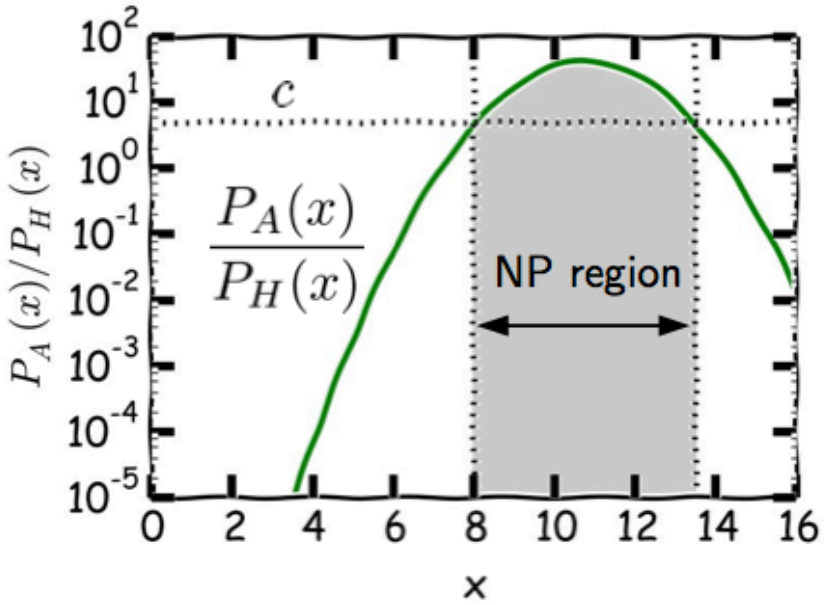
$$\frac{P_A(x)}{P_H(x)} > c$$

- Here, *c* is fixed such that the test has the desired significance:

$$\int P_H(x)\mathrm{Hv}\left(\frac{P_A(x)}{P_H(x)} - c\right)\mathrm{d}x = \alpha$$

Heaviside step function





Generalising to many data points: log-likelihood ratio

- Consider a larger data set: $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$

$$P_H(\mathbf{x}) = \prod_i P_H(x_i) \qquad P_A(\mathbf{x}) = \prod_i P_A(x_i)$$

- It is convenient to define the ***log-likelihood ratio***:

$$\Lambda \equiv -2 \ln \frac{P_H(\mathbf{x})}{P_A(\mathbf{x})} = -2 \sum_i \ln \frac{P_H(x_i)}{P_A(x_i)}$$

- The threshold *c* for rejecting the null hypothesis is obtained from:

$$\int_c^\infty P(\Lambda|H) d\Lambda = \alpha$$

Nested composite hypotheses

- Recall from Week 4:
  - ***simple*** hypothesis: no unknown terms to the model
  - ***composite*** hypothesis: one or more unknown terms known as the model's *free parameters*
- We can consider a null hypothesis which is a subset of the alternative hypothesis:
  - Alternative hypothesis**:** composite model with *n* free parameters

$$P(\mathbf{x}|\theta_1, \theta_2, ..., \theta_n)$$

  - Null hypothesis: composite model with *k* free parameters, and n-*k* constraints:

$$\theta_1, \theta_2, ..., \theta_k \quad \text{free}$$

$$\theta_{k+1}, \theta_{k+2}, ..., \theta_n \quad \text{fixed}$$

- The null hypothesis of two ***nested*** composite models typically lies on a submanifold of the parameter space of the alternative model

Reminder: confidence intervals

- Confidence intervals are the ‘error bars’ on an estimator, e.g. a measured mean count rate in a bin, or a best-fitting estimate of a model parameter (or more generally, an MLE).
- The ***coverage*** of a confidence interval gives the fraction of estimates, or ***confidence level***, for which the true value lies within the interval.
- We can use the inverse of the Hessian of the log-likelihood to approximate the variance of the estimator, the square-root of which for a normally distributed log-likelihood, gives the 1-sigma error (~68.4% coverage)

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

- We also saw that a ‘graphical method’ could be used to find the confidence interval more precisely by looking where the log-likelihood drops by 0.5 from the maximum:

$$L(\hat{\theta} \pm \sigma_{\hat{\theta}}) \approx L_{\max} - \frac{1}{2}$$

- But what if the log-likelihood isn’t normally distributed, and/or more precise estimates, different coverages are needed?

Nested hypotheses example: exact error bars – the confidence belt

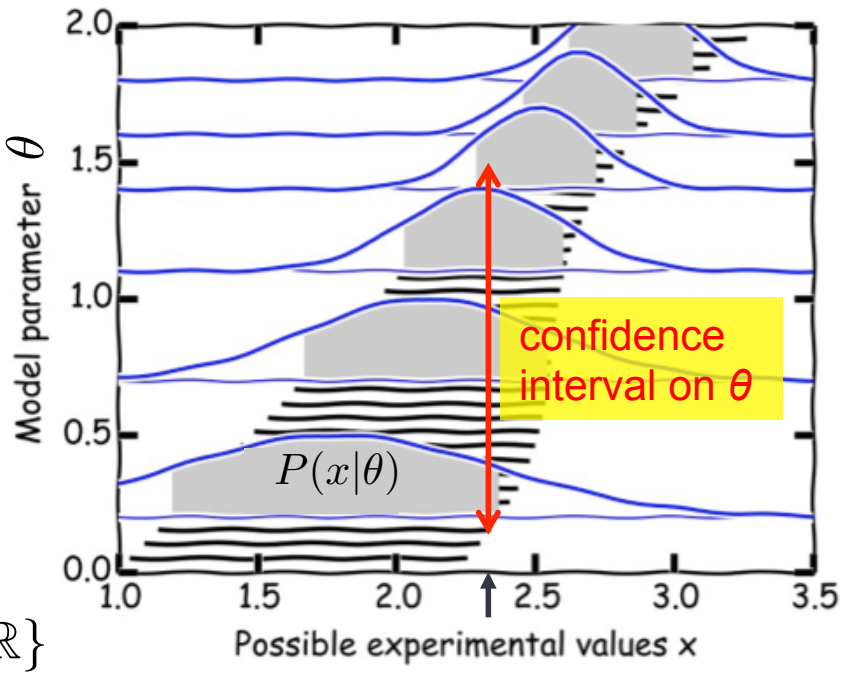
- We can consider confidence intervals as a form of nested composite hypothesis test:
  - *Null hypothesis*: model parameter *θ* fixed to certain value (e.g. the MLE $\hat{\theta}$)
  - *Alternative hypothesis*: model describes data but *θ* is unconstrained
  - *Confidence interval*: all values of *θ* for which the null hypothesis is *not rejected*

e.g. consider a class of hypotheses with one free parameter *θ*. We can define an ***acceptance interval***: $[x_0(\theta), x_1(\theta)]$

where: $$\int_{x_0(\theta)}^{x_1(\theta)} P(x|\theta)\mathrm{d}x = 1 - \alpha$$

This defines the ***confidence belt***
For a given observation x_{obs} the ***confidence interval*** (for coverage 1-*α*) is given by the values of *θ* for which the acceptance interval contains x_{obs}:

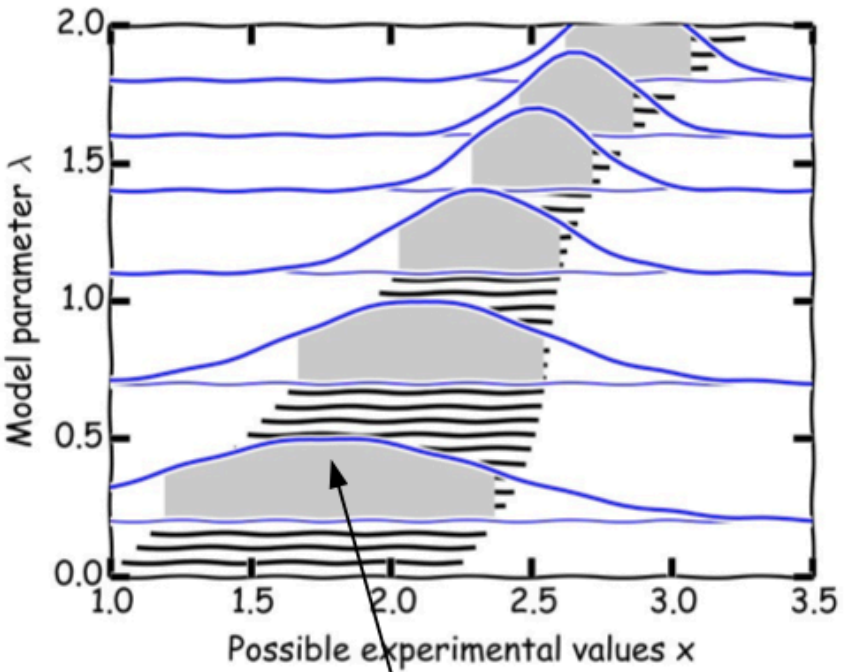
$$I(x_{\mathrm{obs}}) = \{x_0(\theta) \leq x_{\mathrm{obs}} \leq x_1(\theta) | \theta \in \mathbb{R}\}$$

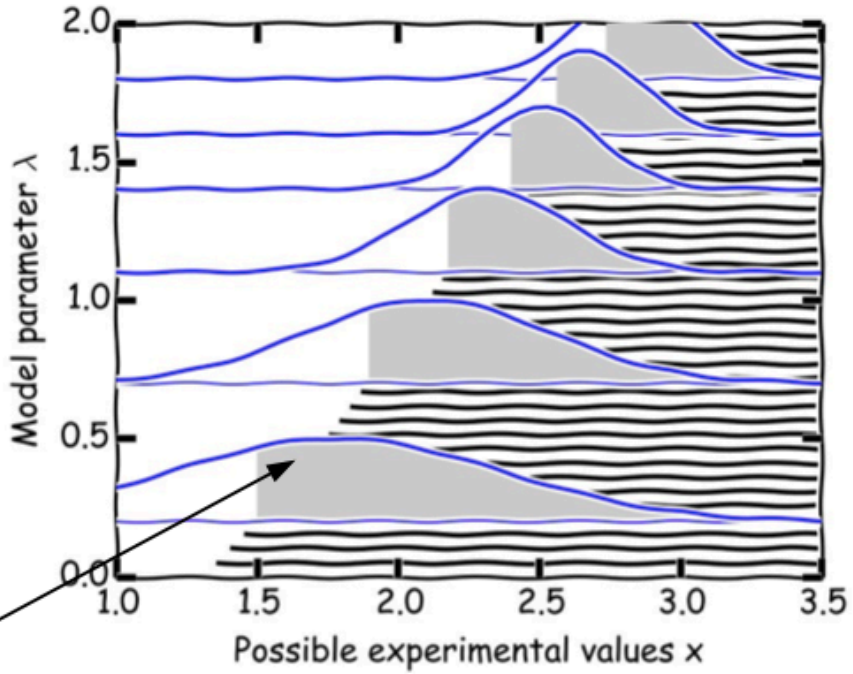


One and two-sided confidence regions

Two-sided: use to define an 'error bar'

One-sided: use to define an upper limit (or lower limit, if one-sided in the other direction!)





The PDFs for a given model parameter λ are identical!

Likelihood ratio construction of confidence regions

We can construct a confidence region using the likelihood ratio approach, if we set the null hypothesis to correspond to a given value of the parameter while the alternative is the parameter MLE, and proceed to search for the region where the likelihood ratio is ***less than*** the critical value for rejection (i.e. we accept the non-MLE value):

$$I(x_{\rm obs}) = \left\{ \underbrace{2\ln \frac{P(x_{\rm obs}|\hat{\theta}_1,\ldots,\hat{\theta}_n)}{P(x_{\rm obs}|\theta_1,\ldots,\theta_k,\hat{\theta}_{k+1},\ldots,\hat{\theta}_n)}}_{\equiv \Lambda(x_{\rm obs},\theta_1,\ldots,\theta_k)} < c(\vec{\theta}) \,|\, \vec{\theta} \in \mathbb{R}^k \right\} \qquad \hat{\theta}_i : \text{ MLE}$$

such that:

$$\int P(\vec{x}|\theta_1,\ldots,\theta_n)\theta_{\rm H}\,(\underbrace{\tilde{c}(\theta_1,\ldots,\theta_n)}_{\simeq c(\theta_1,\ldots,\theta_k)} - \Lambda(x,\theta_1,\ldots,\theta_k))\,dx = \alpha$$

In principle this can be difficult to calculate, but in the large sample limit we can apply a handy theorem…

The distribution of log-likelihood ratio: Wilks' theorem

- Consider the log-likelihood ratio for an alternative hypothesis of a model with *n* free parameters vs. a nested null hypothesis of the same model but with *k* constrained parameters and *n*-*k* free parameters:

$$\Lambda(\theta_1, ..., \theta_k | x) = 2 \ln \frac{l(\hat{\theta}_1, ..., \hat{\theta}_n | x)}{l(\theta_1, ..., \theta_k, \hat{\theta}_{k+1}, ..., \hat{\theta}_n | x)}$$

- If the data *x* is distributed according to the likelihood function *l* for the null hypothesis, then ***in the large sample limit***

$$\Lambda \sim \chi^2_k$$

- i.e. the log likelihood ratio is distributed as chi-squared with *k* degrees of freedom.
- Thus we can determine confidence intervals for our MLEs, and more…

Wilks' theorem: simple demonstration

Recall that (week 4), for normally distributed data, the weighted least squares statistic is:

$$X^2(\theta) = -2L(\theta) + const$$

And that this is distributed as:

$$p(X^2_{\min}|H_0) \sim \chi^2(\nu)$$

where ν is the number of degrees of freedom. For *n* data points and *m* free parameters in the model: $\nu = n - m$

Thus it is easy to see that the log-likelihood ratio for a model with *k* additional free parameters follows a distribution of variable distributed as χ^2_ν minus a variable distributed as $\chi^2_{\nu-k}$

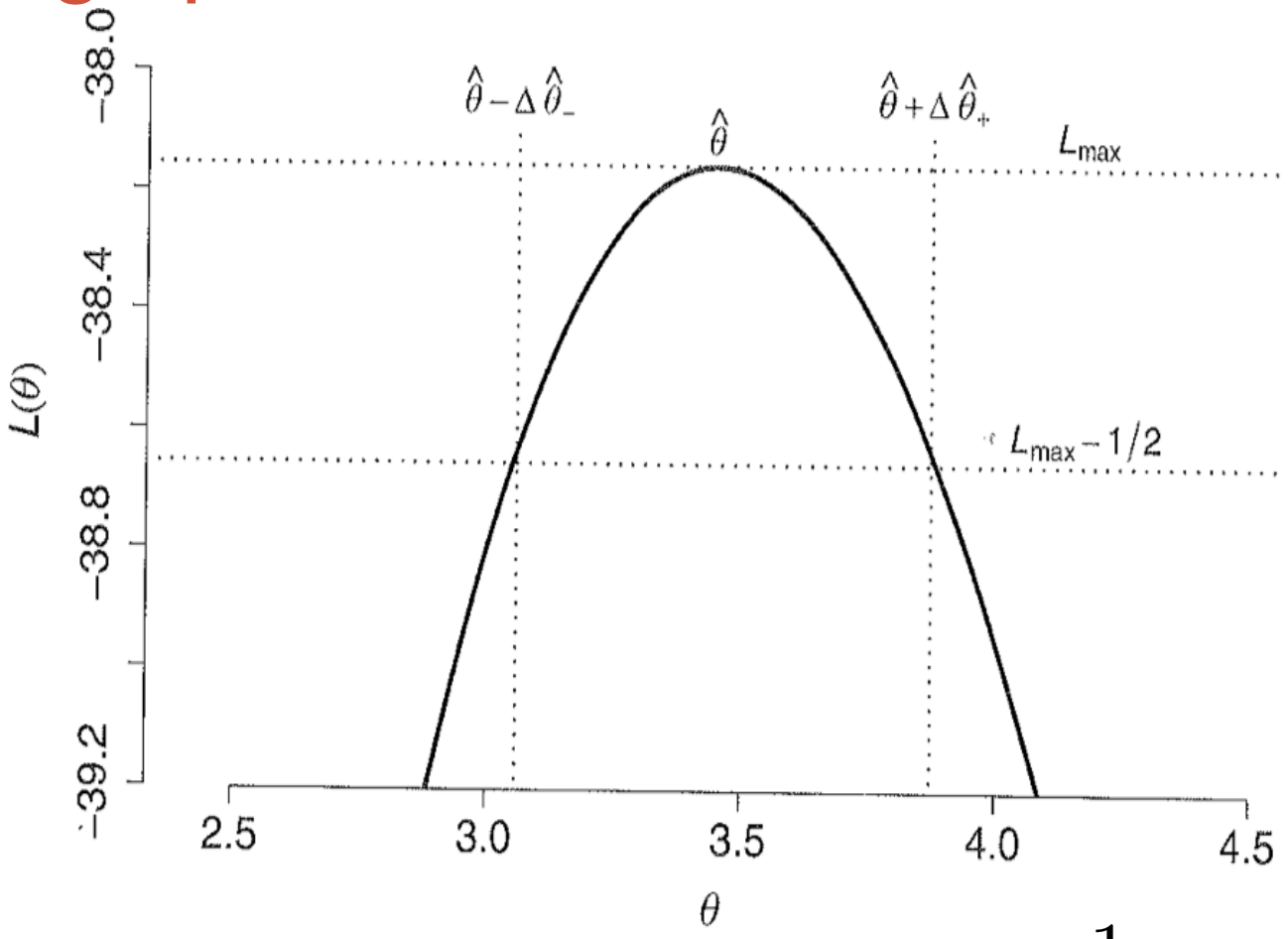
Remembering that a chi-squared distribution arises from a sum of ν squared standard normals, it is easy to see in this case that $\Lambda \sim \chi^2_k$

Wilks' theorem and likelihood ratio: general applicability

- It is important to stress that Wilks' theorem doesn't explicitly require chi-square distributed log-likelihoods. Instead it holds in the much more general case that the ratio is chi-square distributed.
- This requires that the MLEs being kept fixed in the null are normally distributed. The data or other free MLEs can have very non-normal distributions (e.g. this can be used for non-binned data).
- Other methods you will encounter, e.g. *F*-test, Pearson's chi-squared test are either forms of likelihood ratio test or approximations thereof, the likelihood ratio is therefore the most general and widely used.

- However, the use of likelihood ratios assumes the hypotheses being compared are ***nested*** (i.e. the null is a subset of the alternative model). This may not be satisified by your model comparison, in which case alternatives (e.g. relative likelihood) must be looked at (beyond this course)

Application to confidence intervals: why the graphical method works:



$$L(\hat{\theta} \pm \sigma_{\hat{\theta}}) \approx L_{\max} - \frac{1}{2}$$

Application to general hypothesis testing

- We can use Wilks' theorem to test the hypothesis that a particular model parameter(s) can be fixed at a given value or left free when comparing two data sets.
- Many examples, e.g.:
  - Compare a known mean with a measured (fitted) mean
  - Compare means or variances of two data sets: are they consistent with the same values or with different values?
  - Fit a model to data, does one or more of the model parameters need to be different from some pre-set values (e.g. predicted by aspects of the model, or prior data)?
  - Fit a model to data, how far should I go in freeing up model parameters to fit the data, when should I stop (i.e. when am I 'overfitting' the data)?
  - Compare the model parameters from different data sets (e.g. model fits to spectra): do some parameters change between data sets and do some remain the same?
- More details of how this works in this week's tutorial.