Lecture Recording

- **Note: These lectures will be recorded and posted onto the IMPRS website**

- Dear participants,

- We will record all lectures on "*Making sense of data: introduction to statistics for gravitational wave astronomy*", including possible Q&A after the presentation, and we will make the recordings publicly available on the IMPRS lecture website at:

    - https://imprs-gw-lectures.aei.mpg.de/2023-making-sense-of-data/

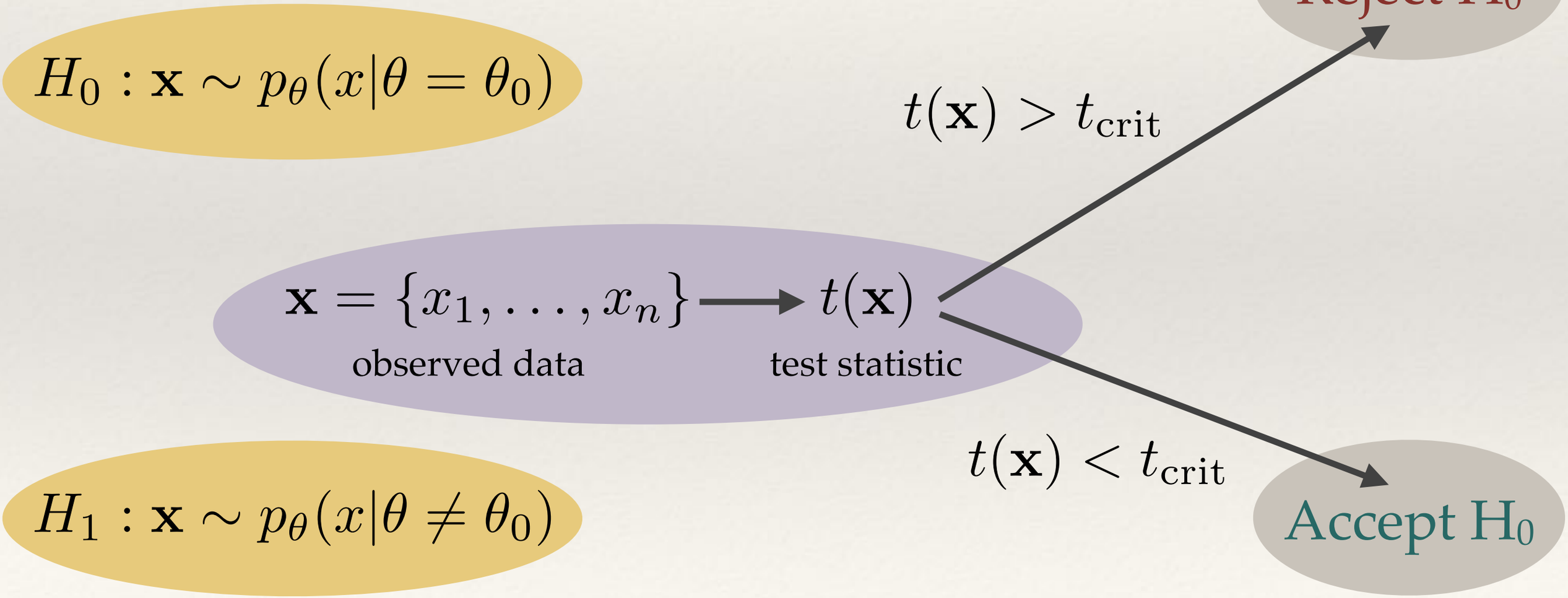
- By participating in this Zoom meeting, you are giving your explicit consent to the recording of the lecture and the publication of the recording on the course website.

Making sense of data: introduction to statistics for gravitational wave astronomy

Lecture 3: hypothesis testing

AEI IMPRS Lecture Course

Jonathan Gair jgair@aei.mpg.de



Hypothesis testing: key concepts

- Having observed data **x** we often want to ask if it is consistent with some pre-conceived assumptions, for example the form of the probability distribution from which the data is drawn or the parameters of that distribution.

- **Hypothesis testing** is usually formulated as a test of a reference **null hypothesis**, H_0, against an **alternative hypothesis**, H_1.

- If a hypothesis is completely specified it is called **simple** otherwise it is **composite**.

- **Examples**:

  - H_0: "the average number of gravitational wave events $\{n_1, \ldots, n_7\}$ observed on different days of the week is the same" is **simple**.

  - H_0: "a trigger in a gravitational wave detector is due to noise" is **composite**, as the instrumental noise distribution is not completely specified.

  - H_0: "the number of gravitational wave events per year is Poisson(λ)" is **composite**.

Hypothesis testing: key concepts

- The outcome of a hypothesis is a decision to **reject** or **accept** (**not to reject**) the null hypothesis.
- The decision is based on the value of a **test statistic**, $t(x)$. Values of the test statistic leading to acceptance of the hypothesis form the **acceptance region**. Values leading to rejection form the **critical region** (or **rejection region**).
- There are two types of error that can be made
  - Reject H_0 when H_0 is true - Type I error
  - Fail to reject H_0 when it is false - Type II error
- The probability of a Type I error, α, is the **significance level** (or **size**) of the test.
- 1 - the probability of a Type II error, $\eta = 1 - \beta$, is the **power** of the test. This is the probability of correctly rejecting H_0.
- Type I errors are considered worse, so we usually quote the significance when describing test results or comparing tests.

Test statistics

- Test statistics used for hypothesis testing need to have certain properties

Definition 10. *A real-valued function $t(\mathbf{x})$ on $\mathcal{X}$ is a test statistic for testing H_0 iff*

(i) values of t are **ordered** *with respect to the evidence for departure from H_0*

(ii) the distribution of $T = t(\mathbf{X})$ under H_0 is known, at least approximately. For composite H_0 the distribution should be (approximately) the same for all simple hypotheses making up H_0.

- In traditional hypothesis testing a threshold is set on the test statistic and values exceeding that threshold lead to rejection.

- It is now common to quote the **p-value** or **significance probability** of a test result. This is the smallest significance level at which the observed test statistic would have led to rejection of the hypothesis.

$$p = \mathbb{P}(T \geq t(\mathbf{x}) | \, H_0)$$

Alternative hypothesis

- The alternative hypothesis can be left **unspecified**, leading to a **pure significance test**. This avoids having to specify H_1.
- The choice of test statistic can be based on the type of deviation from H_0 that the tester is interested in, e.g., look for clustering in observed right ascensions of gravitational wave sources.
- **Goodness of fit** tests compare the sample distribution function, or histogram of event frequencies to the null distribution.
- **Example**: event frequencies on days of the week. Use **Pearson's chi-squared test**, comparing

$$X^2 = \sum_{i=1}^{7} \frac{\left(x_i - \frac{n}{7}\right)^2}{\frac{n}{7}} \quad \text{with} \quad \chi^2_6$$

- Alternative hypotheses can also be **specified**, e.g.

$$H_1 : \theta \in \Theta_1 \subset \Theta \setminus \{\theta_0\}$$

Critical regions

- Tests can also be defined in terms of **critical regions** instead of test statistics.

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size** α if

$$\mathbb{P}(\mathbf{X} \in R_\alpha | \, H_0) = \alpha$$

- Points in R_α are regarded as inconsistent with H_0 or "significant at level α".

- A **significance test** is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \quad \text{if } \alpha_1 < \alpha_2$$

- The **significance probability** (**p-value**) for data **x** is

$$P = \inf(\alpha; \mathbf{x} \in R_\alpha)$$

- Tests based on test statistics have critical regions of the form

$$R_\alpha^t = \{\mathbf{x} : t(\mathbf{x}) \geq t_\alpha\} \qquad \mathbb{P}(\mathbf{X} \in R_\alpha^t | \, H_0) = \mathbb{P}(t(X) \geq t_\alpha | \, H_0) = \alpha$$

Confidence intervals from critical regions

- Critical regions for hypothesis tests provide another way to obtain confidence intervals. Suppose $R_\alpha(\psi_0)$ denotes a size-α critical region for testing

$$H_0 : \psi = \psi_0 \qquad \text{versus} \qquad H_1 : \psi \neq \psi_0$$

- Define

$$S_\alpha(\mathbf{x}) = \{\psi_0 : \mathbf{x} \notin R_\alpha(\psi_0)\}$$

- then $\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi_0; \psi_0, \lambda) = \mathbb{P}(\mathbf{X} \notin R_\alpha(\psi_0) : \psi_0, \lambda) = 1 - \alpha \quad \forall \psi_0, \lambda$
- so $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ confidence interval for ψ.
- **Example**: For n IID exponential random variables the best size-α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$ is

$$R_\alpha(\lambda_0) = \left\{ \mathbf{x} : \sum x_j > \frac{1}{2\lambda_0} \chi^2_{2n}(\alpha) \right\}$$

- which leads to the $(1-\alpha)$ confidence region

$$\left\{ \lambda_0 : \lambda_0 < \frac{1}{2\sum x_j} \chi^2_{2n}(\alpha) \right\}$$

Hypothesis test examples: z-test

- We observe data

$$X_1, \ldots, X_n \sim N(\mu_1, \sigma^2), \qquad Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$$

- We assume σ^2 is known and want to test

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

- The statistic

$$Z = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\sigma}$$

- follows an *N(0,1)* distribution under H_0 and so the critical region takes the form

$$|z| > z_{\frac{\alpha}{2}} \qquad \mathbb{P}(X \sim N(0,1) > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

Hypothesis test examples: t-test

- As in the previous example, we observe data

$$X_1, \ldots, X_n \sim N(\mu_1, \sigma^2), \qquad Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$$

- We now assume σ^2 is unknown and want to test

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

- The statistic

$$T = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\hat{\sigma}} \qquad \text{where } \hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_i - \bar{Y})^2\right)$$

- follows an t_{m+n-2} distribution and so the critical region takes the form

$$|t| > t_{\frac{\alpha}{2}}$$

Hypothesis test examples: F-test

- We observe n_i samples, denoted x_{ij} for $j=1,\ldots,n_i$, in each of k categories and we want to test the hypothesis that the means in all the families are equal. We denote the sample mean in each group by

$$\bar{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

- and the overall sample mean by

$$\bar{X}_{\bullet\bullet} = \frac{1}{N}\sum_{ij} X_{ij}, \qquad N = \sum_{i=1}^{k} n_i$$

- We define the **between samples sum of squares** and **within samples sum of squares** by

$$SS_b = \sum_i n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 \qquad SS_w = \sum_{ij}(x_{ij} - x_{i\bullet})^2$$

- The test statistic $F = \dfrac{(N-k)SS_b}{(k-1)SS_w}$ follows an $F_{k-1,n-k}$ distribution.

- Critical regions take the form

$$F > F_{k-1,N-k}(\alpha)$$

Calculating test statistic thresholds

- Thresholds for hypothesis tests can be constructed in three ways

  - **Analytically**: the distribution of the test statistic may take a known form, e.g., testing need for parameters in a linear model

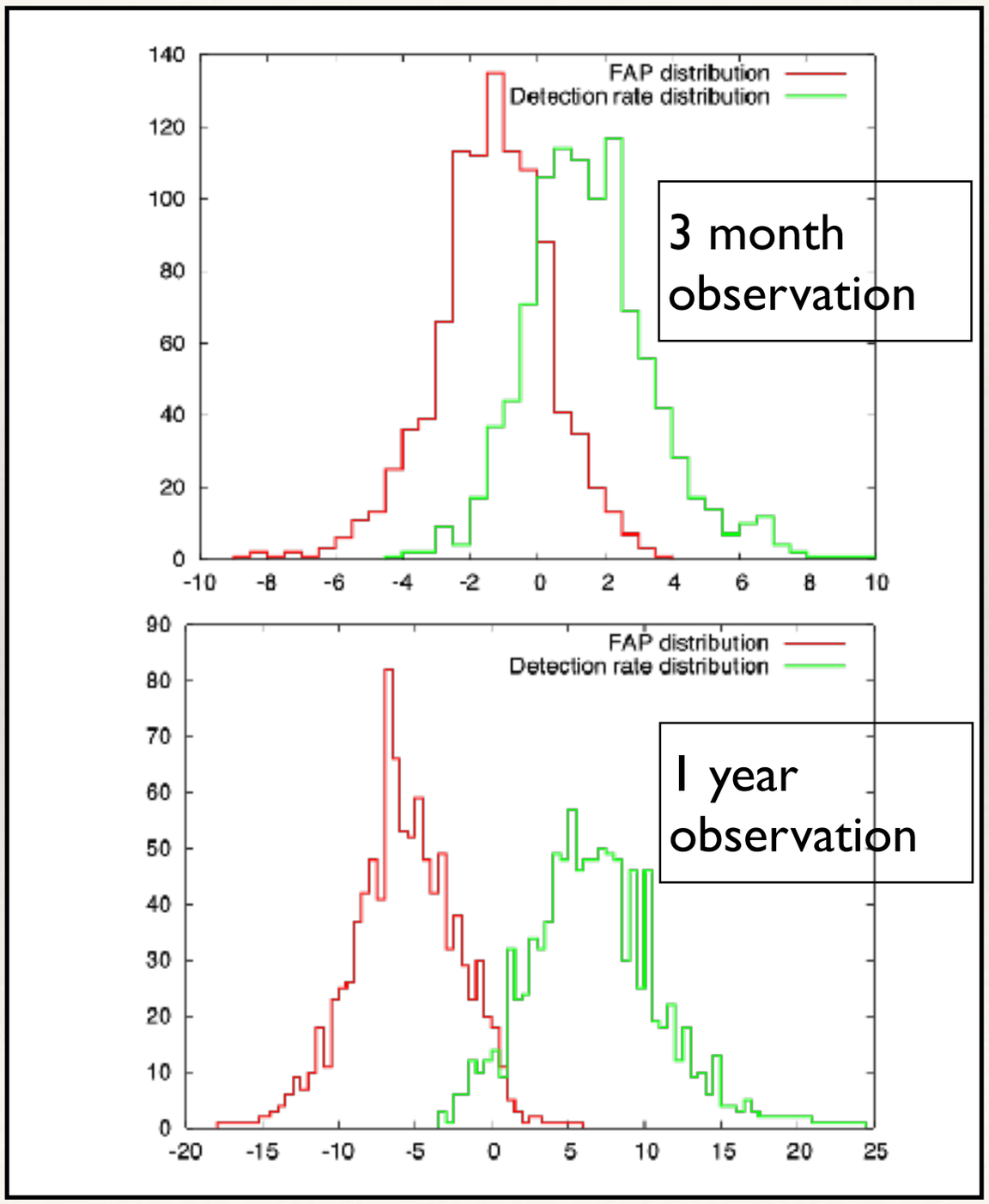
$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2} \quad \text{for testing } \beta_1 = 0$$

  - **Using a Normal approximation**: depending on the form of the test statistic, the CLT can be used to approximate the distribution, e.g.,

$$X = \sum x_j \sim N\left(\frac{n}{\lambda_0}, \frac{n}{\lambda_0^2}\right) \quad \text{for testing } \lambda = \lambda_0 \text{ in } \mathcal{E}(\lambda_0)$$

  - **From a simulation study**: H_0 is normally fully specified, so it can be used to numerically construct the distribution of $t(\boldsymbol{x})$.

- The power of the test can be similarly evaluated.



Caution: multiple testing corrections

- Often the same data will be used for multiple hypothesis tests. If m independent tests of significance α are carried out on the same data, the combined significance is

$$1 - (1 - \alpha)^m = \alpha_c$$

- To achieve a certain target significant for the set of tests, the individual tests should have significance $\alpha = 1 - (1 - \alpha_c)^{1/m}$

- For small significances and numbers of tests, this is approximately $\alpha \approx \alpha_c/m$, which is called the **Bonferroni correction**.

- The total significance can be divided unevenly between the different tests. The **Holm-Bonferroni method** sets $\alpha_i = \alpha_c/(m - i + 1)$, where i labels the tests in order of p-value (starting from the smallest).

- Multiple tests are usually not really independent, so these are all conservative procedures. The true significance of the family of tests must usually be evaluated through simulation.

- In LIGO this effect is referred to as a **trials factor**.

Caution: don't change the question!

- Hypothesis tests may be more or less specific based on prior information. Avoid the temptation to make them more specific **after observing the data**.

- **Example**: LIGO observes for 8 months from January to August and sees (1, 0, 0, 0, 0, 1, 1, 4) events. Is the excess of events in August significant?

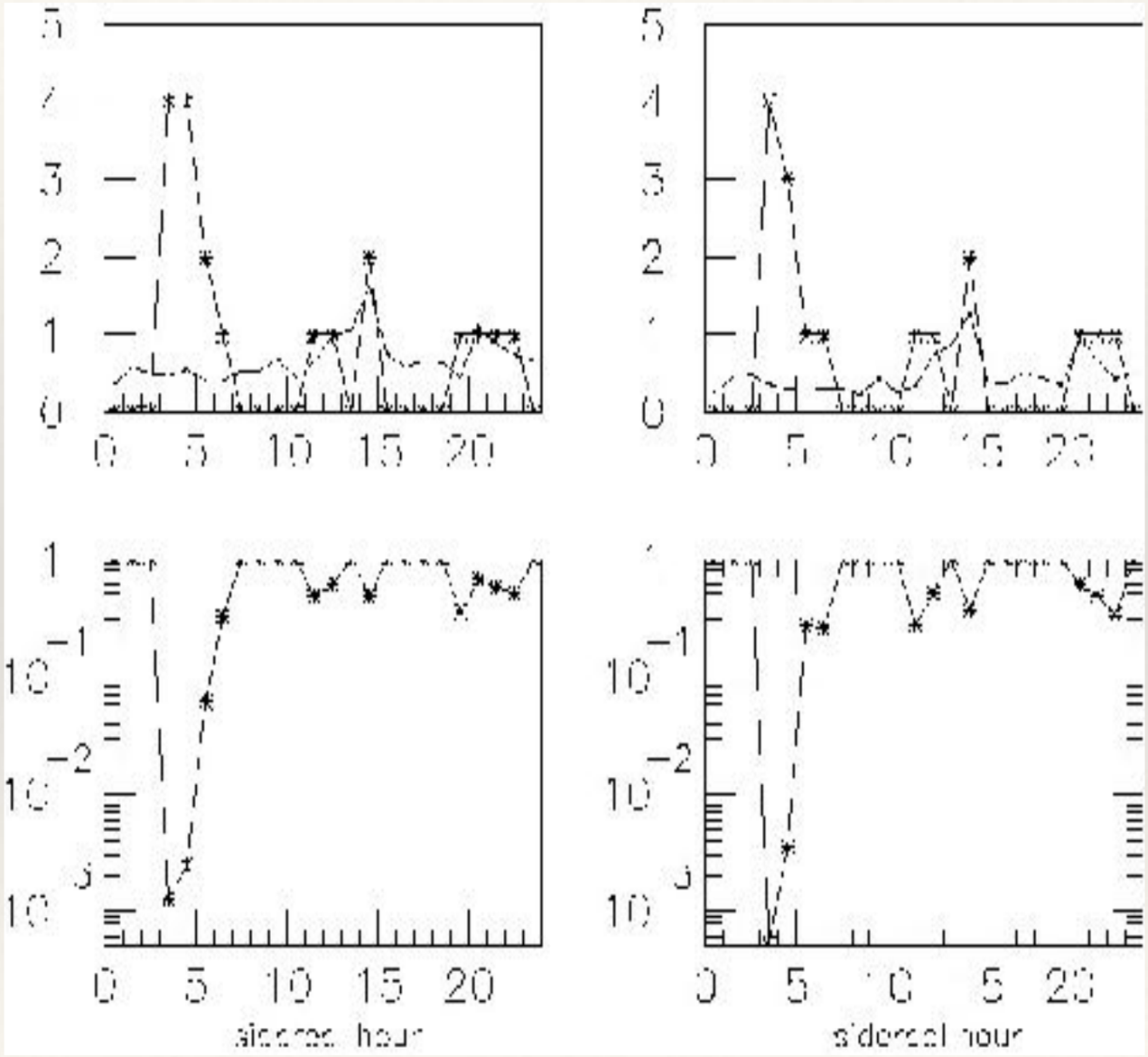
  - The probability of seeing 4 or more events in a specific month is ~1.2% (assuming a Poisson distribution with rate 0.875) or ~0.62% (assuming a multinomial distribution with equal probabilities in all bins and 7 events).

  - The correct question is "How improbable is it to see 4 or more events in one month out of the eight?". The probabilities are then 8 times higher, giving 9.8% or 5% respectively.

  - We can use past data to inform future tests, but these aren't necessarily more powerful than analysing the combined data set. Suppose the next set of observations is (0, 1, 0, 1, 1, 0, 0, 2) then seeing 2 events in August has a 12% probability (in the multinomial analysis). But the combined observations of (1, 1, 0, 1, 1, 1, 1, 6) have probability of 0.18%.

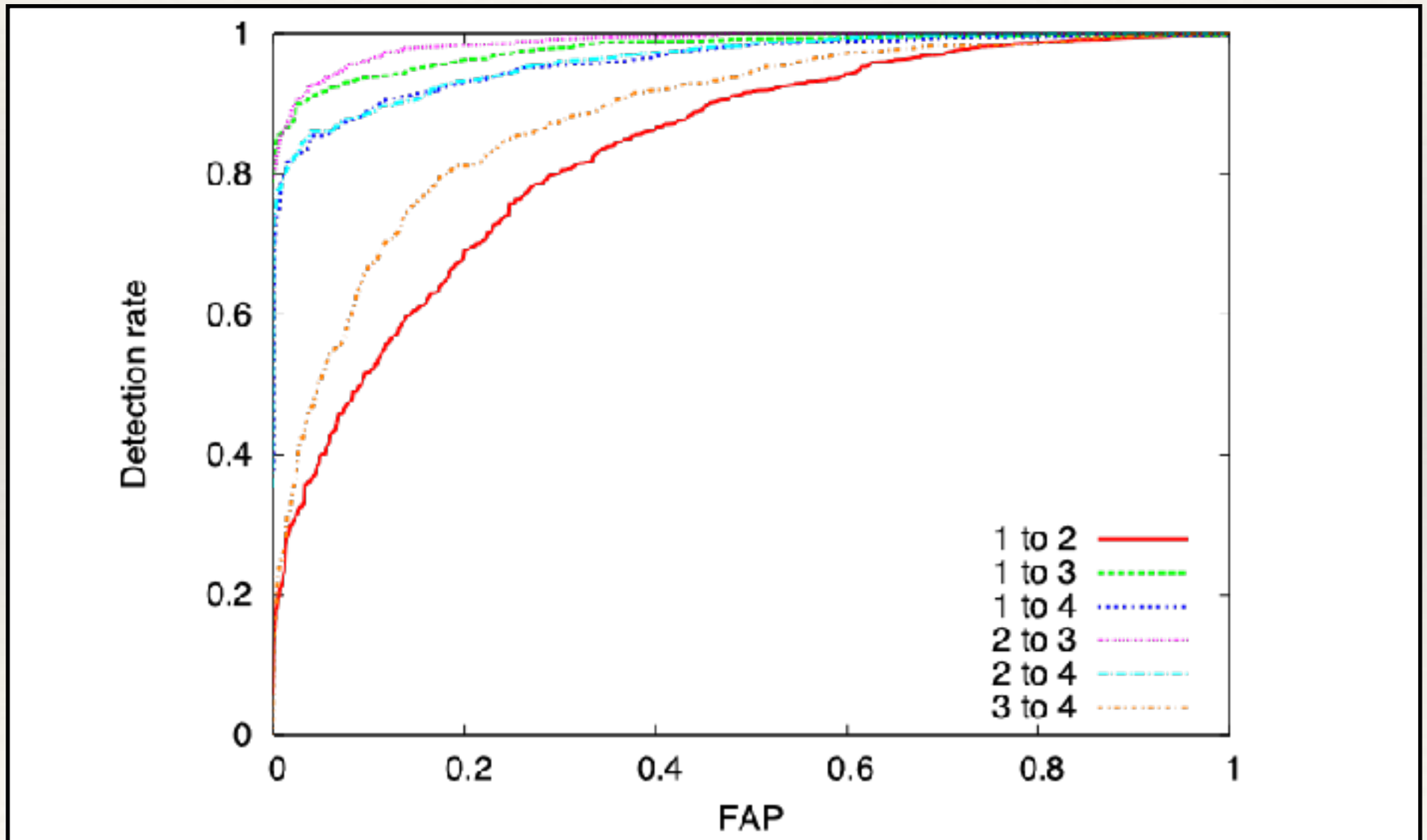
Caution: don't change the question!

- In 2002, the EXPLORER and NAUTILUS teams announced an excess of events towards the galactic centre, based on an excess of events in one bin.
- After seeing the data and realising that bin corresponded to increased sensitivity toward the galactic centre, they decided that they should ask "is there an excess in this particular bin?".
- Such an excess in one (unspecified) bin was not significant.
- The observation was not reproduced in subsequent data.



ROC Curves

- A **receiver operator characteristic** (ROC) **curve** is a plot of the power (or **detection rate**) versus significance (or **false alarm probability**).
- Tests with ROC curves that are **further from the diagonal** are better, i.e., more powerful.



3 month LISA mission, 1 interferometer, SNR threshold 20.

Designing tests: Neyman-Pearson Lemma

- The "best" test is the **most powerful** test at a given significance. Under certain circumstances the best test is given by the **likelihood ratio**

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \frac{L(\theta; \mathrm{H}_1)}{L(\theta; \mathrm{H}_0)}$$

- For testing a simple hypothesis against a simple alternative

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1$$

- the **Neyman-Pearson lemma** states that the **optimal** test is a **likelihood ratio test** with critical regions of the form

$$\{\mathbf{x} : r(\mathbf{x}) \geq k_\alpha\} \text{ or } \{\mathbf{x} : \frac{L(\theta;\mathrm{H}_1)}{L(\theta;\mathrm{H}_0)} \geq k_\alpha\}$$

- **Example**: $X_1,\ldots,X_n$ IID from $\mathcal{E}(\lambda)$. H_0 is $\lambda = \lambda_0$ versus H_1: $\lambda = \lambda_1 < \lambda_0$. The optimal test is based on

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\{(\lambda_0 - \lambda_1)\sum x_j\}$$

- with critical regions $\{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$

Designing tests: UMP tests

- If the null or alternative hypotheses (or both) are not simple, the Neyman-Pearson lemma does not apply. What we are instead interested in are **uniformly most powerful** (UMP) **tests**.

Definition 11. *A* **uniformly most powerful** *or UMP test,* $\phi_0(\mathbf{X})$*, of size* α *is a test* $t(\mathbf{x})$ *for which*

(i) $\mathbb{E}_\theta \phi_0(\mathbf{X}) \leq \alpha \quad \forall\, \theta \in \Theta_0$;

(ii) given any other test $\phi(\cdot)$ *for which* $\mathbb{E}_\theta \phi(\mathbf{X}) \leq \alpha \quad \forall \quad \theta \in \Theta_0$*, we have* $\mathbb{E}_\theta \phi_0(\mathbf{X}) \geq \mathbb{E}_\theta \phi(\mathbf{X}) \quad \forall\ \theta \in \Theta_1$.

- The existence of such tests requires that the Neyman-Pearson test takes the same form for all parameter values in the alternative hypothesis, so this in general is not the case.

- However, for one sided testing problems with simple null hypotheses UMP tests exist for any distributions which have **monotone likelihood ratio**.

Designing tests: UMP tests

Definition 12. *The family of densities* $\{p(\mathbf{x}|\theta), \theta \in \Omega_\theta \subseteq \mathbb{R}\}$ *with real scalar parameter* θ *is said to be of* **monotone likelihood ratio** *if there exists a function* $s(\mathbf{x})$ *such that the likelihood ratio*

$$\frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}$$

is a non-decreasing function of $s(\mathbf{x})$ *whenever* $\theta_1 < \theta_2$.

Theorem 5. *Suppose* $\mathbf{X}$ *has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic* $s(\mathbf{X})$ *and we wish to test* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta > \theta_0$, *then a UMP test exists with critical region of the form* $s \geq s_\alpha$.

Corollary 2. *If* $X_1, \ldots, X_n$ *are i.i.d with p.d.f. of the form*

$$p(x|\,\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ *a scalar parameter and* $b(\theta)$ *strictly increasing, then for testing the null hypothesis that* $\theta = \theta_0$ *against* $\theta > \theta_0$ *the LR test has critical regions corresponding to large values of* $s = \sum a(x_j)$ *and is UMP.*

Designing tests: composite hypotheses

- For two-sided tests of the form

$$H_0 : \theta \in [\theta_1, \theta_2] \qquad \text{versus} \qquad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2$$

- UMP tests do not usually exist. However, **uniformly most powerful unbiased** (UMPU) tests may exist.

Definition 13. *A test $\phi(\mathbf{y})$ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called* **unbiased of size** α *if*

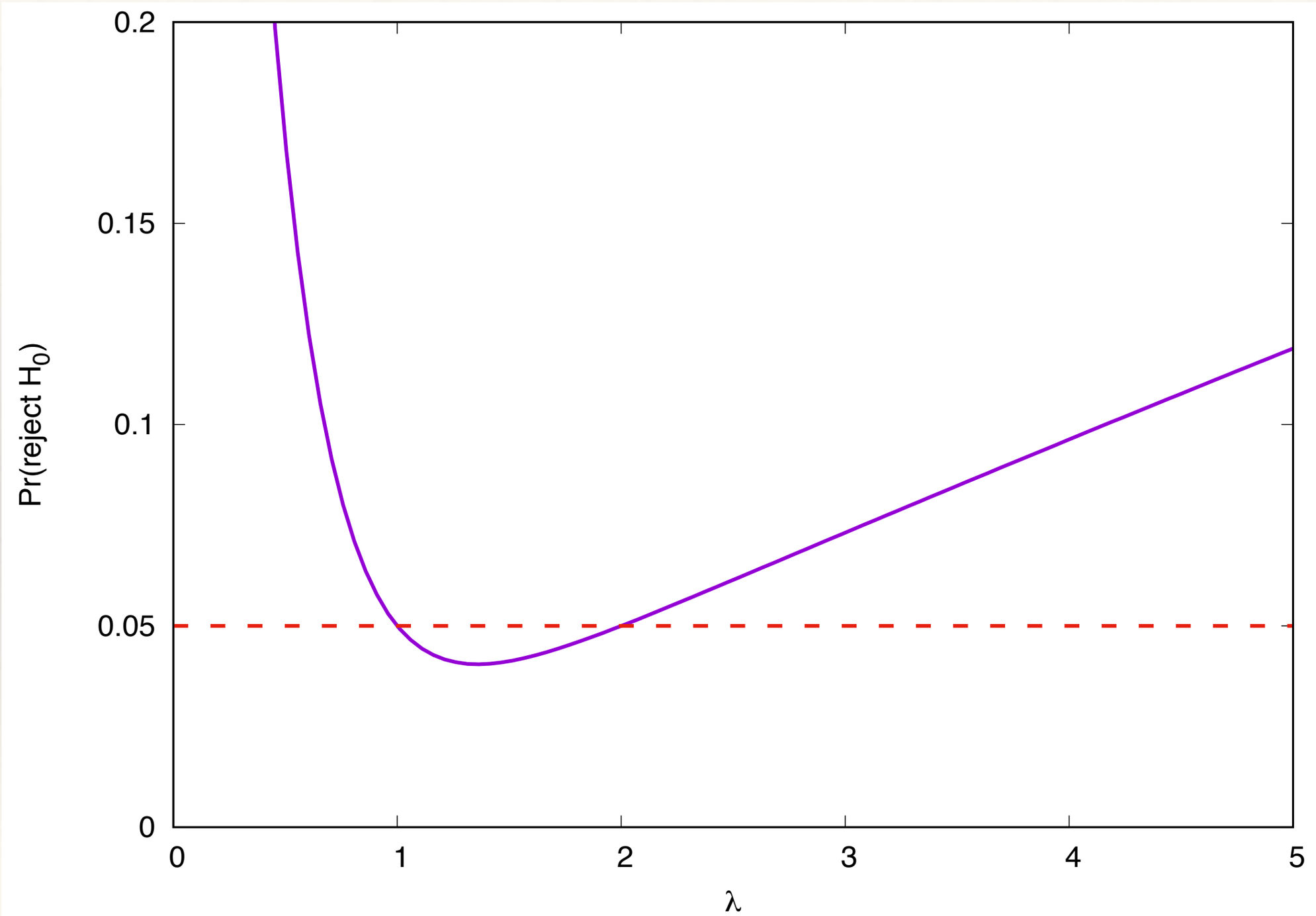
$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \{\phi(\mathbf{Y})\} \leq \alpha$$

and

$$\mathbb{E}_\theta \{\phi(\mathbf{Y})\} \geq \alpha \text{ for all } \theta \in \Theta_1.$$

Definition 14. *A test which is uniformly most powerful among the set of all unbiased tests is called* **uniformly most powerful unbiased**.

Designing tests: UMPU tests



Generalised likelihood ratio test

- If none of the previous results apply, the likelihood ratio is usually still a good test statistic, leading to the **generalised likelihood ratio test**.
- Suppose we are testing

$$H_0 : \vec{\theta} \in \Theta_0 \text{ versus } H_1 : \vec{\theta} \in \Theta_1$$

- We denote by p the difference in the number of degrees of freedom in the two hypotheses, $p = |\Theta_1 - \Theta_0|$, and denote the likelihood ratio by

$$L_X(H_0, H_1) = \frac{\sup_{\vec{\theta} \in \Theta_1} p(x|\theta)}{\sup_{\vec{\theta} \in \Theta_0} p(x|\theta)}$$

- Under certain assumptions the asymptotic distribution is $2 \log L_X(H_0, H_1) \sim \chi^2_p$ and critical regions of the form $2 \log L_X > \chi^2_p(\alpha)$ give tests of approximately size α.