# TESTS, MODEL COMPARISON, MODEL CRITICISM

## 1. TESTS

- Each $T(Y,\theta)$, $R(\theta)$ pair in our construction of a confidence region makes up what is known as a **statistical test** of the **null hypothesis** that $\theta$ is the true value of the parameter. The parameter $\alpha$ is known as the **significance level** or **size** of the test.
- So an exact confidence region can always be interpreted as a collection of statistical tests with significance level $\alpha$.
- More generally, we can consider tests of hypotheses that do not consist of a single point in $\Theta$. For such compound hypotheses, DeGroot and Schervish distinguish level, or significance level, and size. The null hypothesis is some set $\Omega_0 \in \Theta$ and the test still takes the form of a statistic $T(Y)$ and rejection region $R$. Only now it is possible that $P[T(Y) \in R \,|\, \theta]$ differs across $\theta$'s in $\Omega_0$. The standard definitions now say that the test has significance level $\alpha$ if $P[T(Y) \in R \,|\, \theta] \leq \alpha$ for all $\theta \in \Omega_0$, and that it has size $\alpha$ if it has level $\gamma$ for all $\gamma \geq \alpha$.

## 2. POWER

- The **power function** of a test is $P[T(Y) \in R \,|\, \theta]$ considered as a function of $\theta$ over $\Theta \ominus \Omega_0$. (It could be extended to range over $\Omega_0$ also.)
- We would like a test to have a small size and have large values of the power function for $\theta$ outside $\Omega_0$.
- We would like a test to be **unbiased**, meaning that the infimum of $P[T(Y) \in R \,|\, \theta]$ over $\Theta \ominus \Omega_0$ is no smaller than the supremum over $\Omega_0$ of the same thing. (Here $\ominus$ is the set difference operator: $A \ominus B = A \cap B^c$.

## 3. $p$-VALUES

- The $p$-value is generated by considering a family of tests or confidence sets with different significance levels $\alpha$ or confidence levels $1-\alpha$. Usually these are just all the tests generated from a single pivot by varying the size of the test. The $p$-value in the data is then the maximal $\alpha$ at which the hypothesis would be rejected in the sample at hand.
- The $p$-value is a way to summarize results without committing in advance to a test size — the reader can use it to accept or reject based on any size.

## 4. Bayesian analyses that look like tests

- Posterior tail areas. These can be useful descriptions of the shape of the posterior pdf, and they will be close to $p$-values when the likelihood is approximately Gaussian in shape and the parameters are location and scale parameters.
- Often interest centers not on the "null" value of the parameter at which the $p$-value is computed, but at the most likely values. When the likelihood is non-Gaussian in shape, $p$-values can be misleading, since they will not relate to likelihood or posterior pdf shape in the usual way.
- When tests are not based on sufficient statistics, their $p$-values don't connect in a useful way to likelihood. For example, the $\sqrt{\mu}\sum(X_t - \mu)$ pivot in problem 2 in the problem set depends on only one of the two sufficient statistics in the model, and tests based on it do not tell you much about likelihood shape.

## 5

- Posterior probabilities on models. Think of "model number" as a parameter. Integrate other parameters out of the likelihood, leaving a marginal posterior distribution over models.
- This seems much like testing one model as $H_0$ against another model (or other models) as $H_A$. But it turns out to be rather different.
- Example: $X_t \sim$ i.i.d. $N(\mu, 1)$. Consider this model vs. the same model with the restriction $\mu = 0$.
- $\bar{X}$ is a pivot, there is an obvious, standard way to generate a .05 level test (or rather two — two-sided and one-sided).
- To construct posterior probabilities we need prior probabilities $q$ and $1-q$ on the two models. But also a *proper* prior on $\mu$. This is a general fact — meaningful posterior odds on models require proper priors on the model parameters.
- The posterior is proportional to a discrete weight $q\exp(-.5\sum(X_t)^2)$ on the $\mu = 0$ model and a density $(1-q)\int(2\pi)^{-1/2}\exp(-.5\sum(X_t-\mu)^2 - .5\mu^2)\,d\mu$ on the unconstrained model, assuming a $N(0,1)$ prior on $\mu$ in the unrestricted model. Here we are dropping common $(2\pi)^{-T/2}$ factors from both model likelihoods. The $\sqrt{2\pi}$ factor in the second posterior weight comes from the prior on $\mu$.
- Integrating w.r.t. $\mu$, we arrive at a weight on the second model of

$$\frac{e^{-\frac{1}{2}(\sum X_t^2 - (T+1)\bar{X}^2)}}{\sqrt{T+1}},$$

where, because of the prior, We write $\bar{X} = \sum X_t/(T+1)$. The integration w.r.t. $\mu$ is done as usual by completing the square on the quadratic term in the exponent, so that the integrand, as a function of $\mu$, becomes proportional to a Gaussian pdf.

This leads to the log odds ratio

$$\log\left(\frac{q}{1-q}\right) - \frac{1}{2}(T+1)\bar{X}^2 + \frac{1}{2}\log(T+1)$$

Note that with $q = .5$, this criterion favors the unrestricted-$\mu$ model when

$$|\bar{X}| > \sqrt{\frac{\log(T+1)}{T+1}}\,.$$

If instead we favored the restricted model when $|\bar{X}|\sqrt{T} > \Phi\,(.975)$ (a usual 5% test of the $\mu = 0$ null hypothesis), we would reject less often for small $T$, more often for large $T$. The rejection threshold shrinks at the rate $1/\sqrt{T}$ for the standard test, while it shrinks at $\log T/\sqrt{T}$ for the posterior odds criterion. To give you an idea, here is the two-tailed significance level corresponding to the posterior odds ratio of 1 for a few sample sizes:

| Sample size | Significance level |
|---|---|
| 5 | .22 |
| 20 | .088 |
| 80 | .038 |
| 400 | .0144 |
| 4000 | .004 |

The sample size at which the .05 level corresponds to equal posterior odds is about 50.

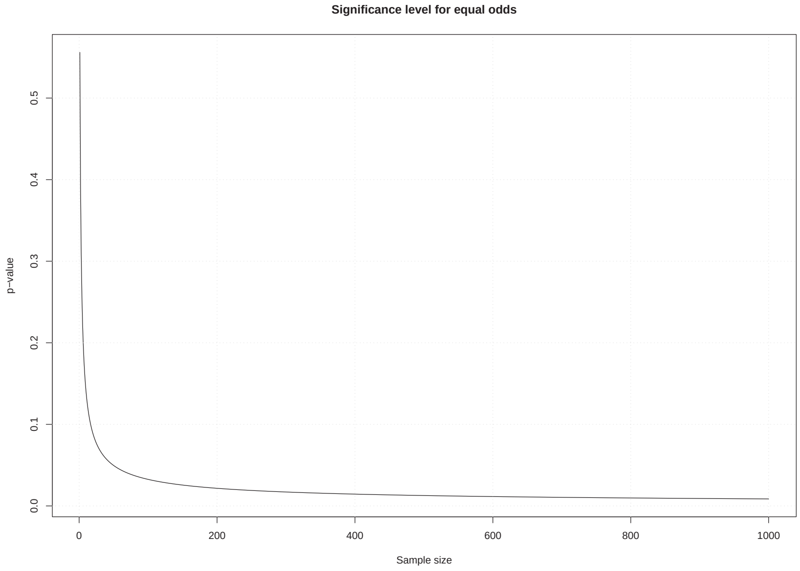
- These results do depend on the prior. If we had taken the prior standard deviation of $\mu$ in the unrestricted model to be 2 rather than 1, the resulting lower pdf in the neighborhood of $\mu = 0$ would have favored the restricted model. Even in a sample of 5, in that case, the equal odds ratio would have corresponded to an 8.8% significance level and at a sample of size 20 to a 2.7% level.

## 6. What are tests and confidence sets for?

(i) Contributions toward characterizing the shape of the likelihood.
(ii) "Is this model, or this restriction on the model, consistent with the data?"
(iii) "Could this apparently interesting result have arisen 'at random'?"
(iv) Which should we use, $H_0$ or $H_A$, for decision-making purposes?

## 7. Bayesian views on tests and confidence sets for these purposes

- Confidence sets that correspond to HPD density regions under a flat prior, like the usual SNLM confidence sets, are indeed useful to describe the likelihood.

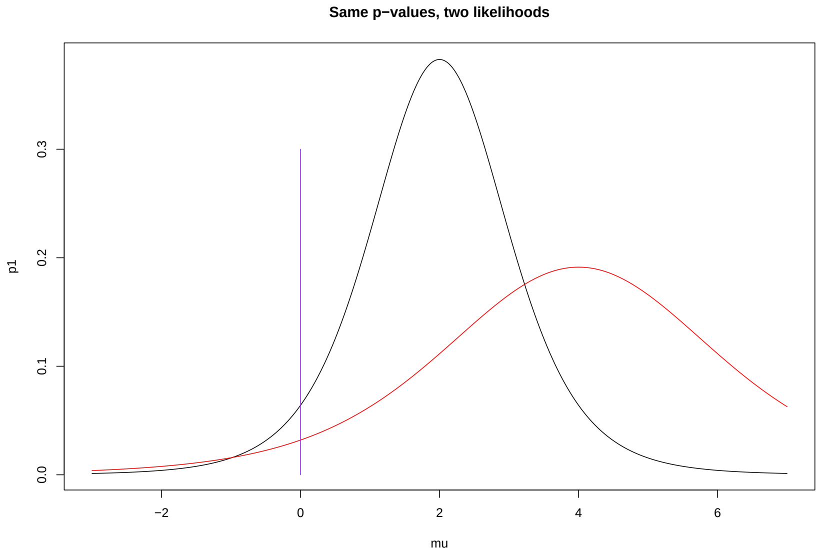


- With point null, point alternative and observation of the test outcome only, test rejection or acceptance, together with the significance level and power of the test, completely characterize the likelihood. Of course usually we can observe more than just the test outcome and in realistic decision problems the point null, point alternative framework is rare.

## 8. Simple example of somewhat instructive $p$-values

$$\{X_1, \dots, X_N\} \sim i.i.d. N(\mu, \sigma^2)$$
$$p - \text{value}: \ 1 - F_{t(N-1)}\left(\frac{\bar{X}}{s}\right) = P\left[t_{N-1} > \frac{\bar{X}}{s}\right] .$$

Here the test family is the family of one-sided tests of $H_0 : \mu = 0$ that reject when $\bar{X}/s$ is too far below zero. The tests are based on a pivot, whose distribution is $t_{N-1}$ under the null. It should be clear that the model, the number of observations, and the $p$-value tell us quite a bit about the likelihood. But they don't fully characterize it. See the diagram. If we add to the number of observations and the $p$-value the point estimate $\bar{X}$, we do have a complete characterization of the likelihood.



## 9. Could the results have arisen "at random"?

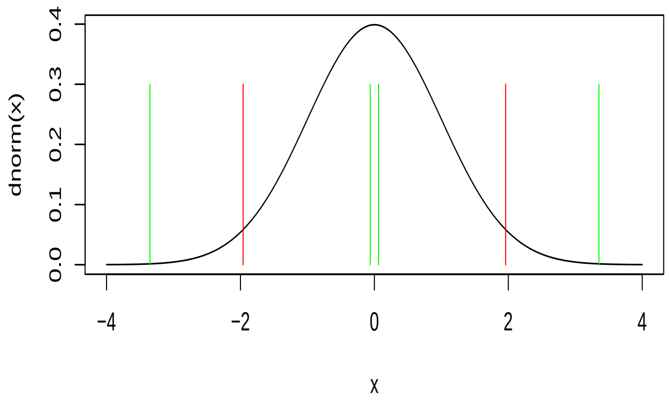
- Bayesians have tended to be annoyed with this question, while some non-Bayesians will say that the ability to answer this question with a "test" is a primary reason for their adherence to a non-Bayesian approach to inference.
- Some Bayesians, including Lancaster and Gelman, Carlin, Stern and Rubin, like the idea of a Bayesian approach to answering this question.
- While some ways of answering the question are sensible, this is true only when there is in the background an implicit class of alternative models. There is really no way, in a Bayesian or non-Bayesian framework, to test a model against no model.

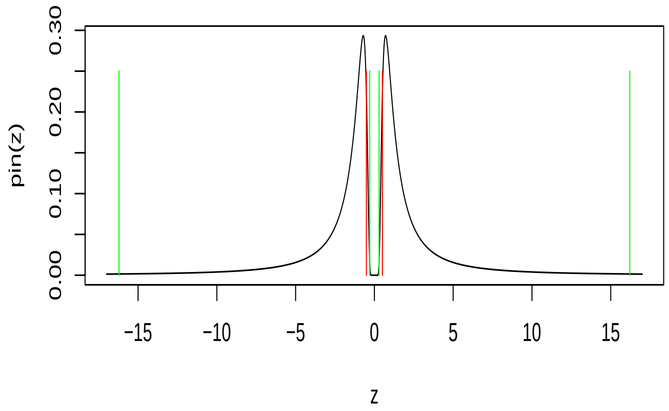
## 10. Example

The sample mean of $X_1, \ldots, X_T$ is $\bar{X}$. Perhaps this seems big, but could a sample mean this big have arisen "at random"? Here again we can calculate $\bar{X}/s$ and check it against the $t_{T-1}$ distribution. If it is far out in the tails we say the result is "significant" — meaning it is unlikely to have arisen "at random". This can be interpreted either as the outcome of an hypothesis test, or as a statement that a flat-prior posterior has little probability so far out in its tail. This situation is shown in the upper panel of the figure, which is a plot of the pdf of $\bar{X}/s$ under $H_0 : \mu = 0$. The red lines on the plot are the boundaries of the acceptance region for $H_0$.

But suppose instead of checking the distribution of $\bar{X}/s$ we check the distribution of $s/\bar{X}$? The lower panel shows this pdf under the null. The rejection region of the upper panel maps into a narrow region around zero in the lower panel. Furthermore, if we had formed a rejection region directly from the lower panel, picking a region of highest pdf values, it would include a narrower band around zero, plus tails of this distribution. The boundaries are marked by the green vertical lines, both here and in the upper panel.

Which rejection region is better? There is no answer without specifying what kind of alternative we have in mind.





## 11. Arbitrariness of density heights

- By monotone transformation of a random variable, we can make the density high or low wherever we like.
- The pdf of $Y = g(X)$ is $f(g^{-1}(Y)))/g'(g^{-1}(Y))$. If $h(\cdot)$ is an arbitrary density function we want to match, we form the corresponding $F_h$ cdf and set

$F(g^{-1}(Y)) = F_h(Y)$, or $g(x) = F_h^{-1}(F(x))$. So long as h was every where non-zero, the inverse function of the cdf $F_h$ will exist.

- So what are the "unusual" or "unexpected" values of $X$ that we could observe to make us reject an assumed distribution for it? Those in low-pdf regions for $X$? Or in low-pdf regions for $g(X)$?
- *Every* real number has zero probability of occurring if $X$ has a positive density everywhere. We form positive-probability rejection regions only by grouping points, and the ways we do this grouping are arbitrary — unless we have an alternative hypothesis in mind.
- If there is an explicit alternative, the sensible thing to do is to base rejection regions on the ratios of the pdf's of the observations under the null and the alternative — that is on the likelihood ratios, which also determine the posterior probabilities.
- Likelihood ratios are invariant under monotone transformations. If $H_0$ gives $X$ the pdf $f(x)$ and $H_A$ gives it the pdf $q(x)$, then the ratio of the two pdf's for $Y = g(X)$ is

$$\frac{f\big(g^{-1}(y)\big)/g'(g^{-1}(y)}{q\big(g^{-1}(y)\big)/g'(g^{-1}(y))} = \frac{f(x)}{q(x)} .$$

## 12. Complications

- Our discussion has applied to a case where the model implies a pdf for the data, so the likelihood across models is determined directly from the observed data.
- Two common complications
  (1) Each model being compared has estimated parameters, so the likelihood varies over these parameters as well as over models.
  (2) Instead of using the likelihood over all the data, we form a test statistic $T(Y)$, which may not be a sufficient statistic, test based on this rather than on $Y$.
- The first means there are many likelihood ratios between models, as we vary the model parameters. One must average them, maximize them, or in some other way aggregate them. Bayesian posterior probabilities average likelihoods using the posterior weights on the parameters. Classical tests that perform well generally do the same thing, or almost the same thing.
- The second means we are not using all information in the sample, if $T(y)$ is not a full list of sufficient statistics. It can happen that likelihood ratios between models depend only on some of, or a few functions of, the sufficient statistics, in which case using only these functions of the data involve no loss. But whether there is a loss, and how big it is, depends on the models being compared. It can be useful to base tests on an easily computed $T(Y)$ even if

there is some loss of information, so long as we know that the information loss is not too great.

## 13. Back to the $\bar{X}/s$ example

- So when would it make sense to reject $H_0 : X_t \sim N(0, \sigma^2)$ on the basis of large values of $\bar{X}/s$?
- The sufficient statistics for the unknown parameter $\sigma$ is $\sum X_t^2$, not $\bar{X}/s$, so it must be that the implicit alternative has a likelihood sensitive to $\bar{X}$ and $s^2$ separately.
- One class of alternatives (the only one?) for which this kind of rejection region does make sense: $H_A : X_t$ i.i.d. $N(\mu, \nu^2)$, with the priors on $\log \nu$ and $\mu$ approximately flat. How do we know this? The Bayesian posterior odds ratios in this case are a function of $\bar{X}/s$.
- Classes of alternatives for which this kind of rejection region makes no sense: $H_A : X_t$ i.i.d. $N(1, \nu^2)$ or $N(\mu, 1)$.
- In this example it is fairly clear intuitively what are the alternatives against which the test is and is not powerful. In more complicated models, tests of a null presented with no discussion of what they are powerful against should be regarded with suspicion. Even in apparently standard cases, it is worthwhile to think through what kinds of alternatives the test would be weak against.
- Often the easiest way to do this is to figure out how the posterior odds ratios depend on the data under various alternatives.

## 14. Bayesian model checking

- Like non-Bayesian "testing against nothing", this attempts to see whether the observed data is "unlikely" given the model, without specifying an alternative model.
- Using a pre-data distribution for the data: Using prior and model, generate, either by simulation or analytically, a distribution for $T(Y)$ that does not condition on the parameter. Look to see if observed $T(Y)$ is toward the center of this distribution.
- Using a post-data distribution: draw from the posterior on the parameters, then for each draw of the parameters, draw a sample $T(Y)$. Again check to see if the observed $T(Y)$ is near the center of the distribution.
- It is hard to see how to make sense of these ideas, as they rely on our being able to distinguish "likely" from "unlikely" values without reference to an explicit alternative hypothesis. As we pointed out above, monotone transformations of the data can make any region of the sample space have arbitrarily high or low density. To reject a model, we should have in mind that

some other model makes the observed data more likely than does the rejected model.

Suppose a farmer has a hypothesis that feeding pigs corn rather than pumpkins makes them grow faster. He starts feeding half his pigs corn, the other half pumpkins. A week later, he looks out the window and sees all his pigs have grown wings and are flying. This is so unlikely, under the model that says pigs grow faster eating corn, that he concludes he must reject the hypothesis that they grow faster eating corn. If this reasoning makes sense to you, you will be interested in further study of checking models without specifying alternatives. See sections 2.4-2.5 of Lancaster.

## 15. Models for discrete data

Most of what we discussed under this heading is in Lancaster's section 5.2. Lancaster shows how to derive the Probit from a choice model, but he doesn't explain the corresponding story for the logit.

In both probit and logit, we can think of the observed variable $y_i$ as equal to 1 if $U_{1i} = x_i\beta_1 + \varepsilon_{1i} > U_{0i} = x_i\beta_0 + \varepsilon_{0i}$, where $\varepsilon_{ji}$, $j = 0, 1$ are independent of each other and of $x_i$ and the data are i.i.d. For the probit model, we assume $\varepsilon_{ji} \sim N(0, 1/\sqrt{2})$, and this leads to the criterion that $y_i = 1$ iff $x_i\gamma > -\nu_i$, where $\nu_i = \varepsilon_{1i} - \varepsilon_{0i}$ is $N(0, 1)$ and $\gamma = \beta_1 - \beta_2$.

For the logit, the question is what distribution of $\varepsilon_{ji}$ in a setup like this would lead to the logit model. The answer is the **log-Weibull** distribution, which has cdf $\exp(-e^{-x})$ and pdf $e^{-x}\exp(-e^{-x})$. This pdf is single-peaked with a max at $x = 0$. It has a more or less bell shape, though its tail drops off much more rapidly to the left than to the right.

Besides logit and probit, you should understand the defects of the linear probability model, which are discussed in Lancaster along the same lines as the lectures.