Estimation, testing, confidence, credibility

Christopher A. Sims
Princeton University
sims@princeton.edu

September 22, 2020

Estimation as special case of decision theory

- If in the standard decision theory setup we make the objective function

$$U(\beta, X, \delta(X)) = F(\beta - \delta(X)) \,,$$

with $F(z)$ maximized at $z = 0$, then we are trying to choose δ to be close to the unknown β. This makes the decision problem an **estimation** problem, and we call $\delta(X)$ an **estimator**.

Quadratic objective

- Often we consider the special case where F is quadratic, i.e.

$$F(\beta, X, \delta(X)) = -(\delta(X) - \beta)' A (\delta(X) - \beta) \, ,$$

  with A positive definite.

- For this quadratic case, if there is no restriction connecting the value of $\delta(x)$ to its value at other values of x, it is optimal to set $\delta(X) = E[\beta \mid X]$.

- Note that this result does not depend on A.

Proof of optimality

If we label $\bar{\beta}(X) = E[\beta \mid X]$, and $\Sigma(X) = \mathrm{Var}(\beta \mid X)$

$$E[-(\delta(X) - \beta(X))'A(\delta(X) - \beta(X)) \mid X]$$
$$= (\delta(X) - \bar{\beta}(X))'A(\delta(X) - \bar{\beta}(X)) + \mathrm{trace}(\Sigma(X)A)\ .$$

The equality above follows from the fact that $E[(\beta - \bar{\beta}(X)) \mid X] = 0$. The **trace** operator takes the sum of all elements of its matrix argument, and has the property that $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.

The univariate normal i.i.d. case

- Last lecture we verified that $\mu \mid \{x_i\}$, with i.i.d. $N(\mu, \sigma^2)$ data and a flat prior, is distributed as a t variate with mean and mode $\bar{x}$, the sample mean. So this is the Bayes estimator for that prior with a quadratic objective function.

- A frequentist approach puts probabilities only on observable data (X) and treats β (or in our $N(\mu, \sigma^2)$ case, (μ, σ^2)) as fixed.

- In this approach, estimators $\delta(X)$ are proposed directly as functions $\delta(X)$ of the data, and the distribution of $F(\beta, X, \delta(X))$ with β fixed and X taken as random is discussed.

Unbiasedness

- In the univariate normal case, if we are interested in estimating μ, a natural proposal for $\delta()$ is the sample mean $\bar{x}$. It is a function of the data with the property that $E[\bar{X} \mid \mu, \sigma^2] = \mu$, regardless of the values of μ and σ^2.

- This makes $\bar{X}$ an **unbiased** estimator of μ.

- Thus in this case, it turns out that $\bar{X} = E[\mu \mid X]$ and at the same time that $E[\bar{X} \mid \mu, \sigma^2] = \mu$.

- This is unusual. Ordinarily, the expected value of a parameter given the data is not an unbiased estimator of the parameter.

Examples where unbiased estimators are clearly suboptimal

- We discussed previously (not on slides) the point that where a parameter has bounded **support**, unbiased estimators can usually be improved at the boundaries of the support. (Support of a distribution: smallest closed set with probability one.)

- Another example, which you'll explore in an exercise: $X_i \sim N(\mu, \mu^2)$. The sample mean is unbiased, but misbehaves when the sample happens to have mean and sample standard deviation far apart.

Testing

- This is a situation where we have a model, or a class of models, that we label "H0", for "**null hypothesis**", that we want to compare to another class of models, "HA" by seeing how well they account for some observed data.

- This is simple and easily understood when neither "H" has unknown parameters, so that we are just choosing between two distributions for the data. This is known as the "point null, point alternative" case.

- The frequentist approach is to propose a function of the data that has the value "Accept H0" or "Reject H0". The properties of the test are then characterized by $P[\text{Reject} \mid \text{H0}]$, known as the "**significance level**" or probability of "**type I error**", and $P[\text{Reject} \mid \text{HA}]$, known as the **power** of the test, or one minus its "**type II error**" probability.

Decision-theoretic view of testing

- We assume our decision must be "Accept" or "Reject", and it can depend on the data, but not on the unknown true distribution.

- We assume our objective function has higher values when we choose Accept and H0 is true or choose Reject when HA is true. Then tests with lower type I and type II errors are better.

- However, in general our choice of test might depend on our prior distribution over H0 and H1 and on how our objective varies when we make errors.

- We have already seen in the rare-disease-testing example how decisions (Do you initiate expensive or painful treatment on a person who has tested positive?, e.g.) could depend strongly on priors, even when type I and type II errors are both small.

You can't test H0 against nothing, or everything

- You might think you can test H0: $X \sim N(0,1)$ without specifying an alternative. Just reject if $|X| > 2$, which has probability close to .05 under the null, so it is a 5% significance level test.

- This test would make sense if HA were a normal distribution with the same variance and a different mean, or a collection of other normal distributions with variance 1 and non-zero mean.

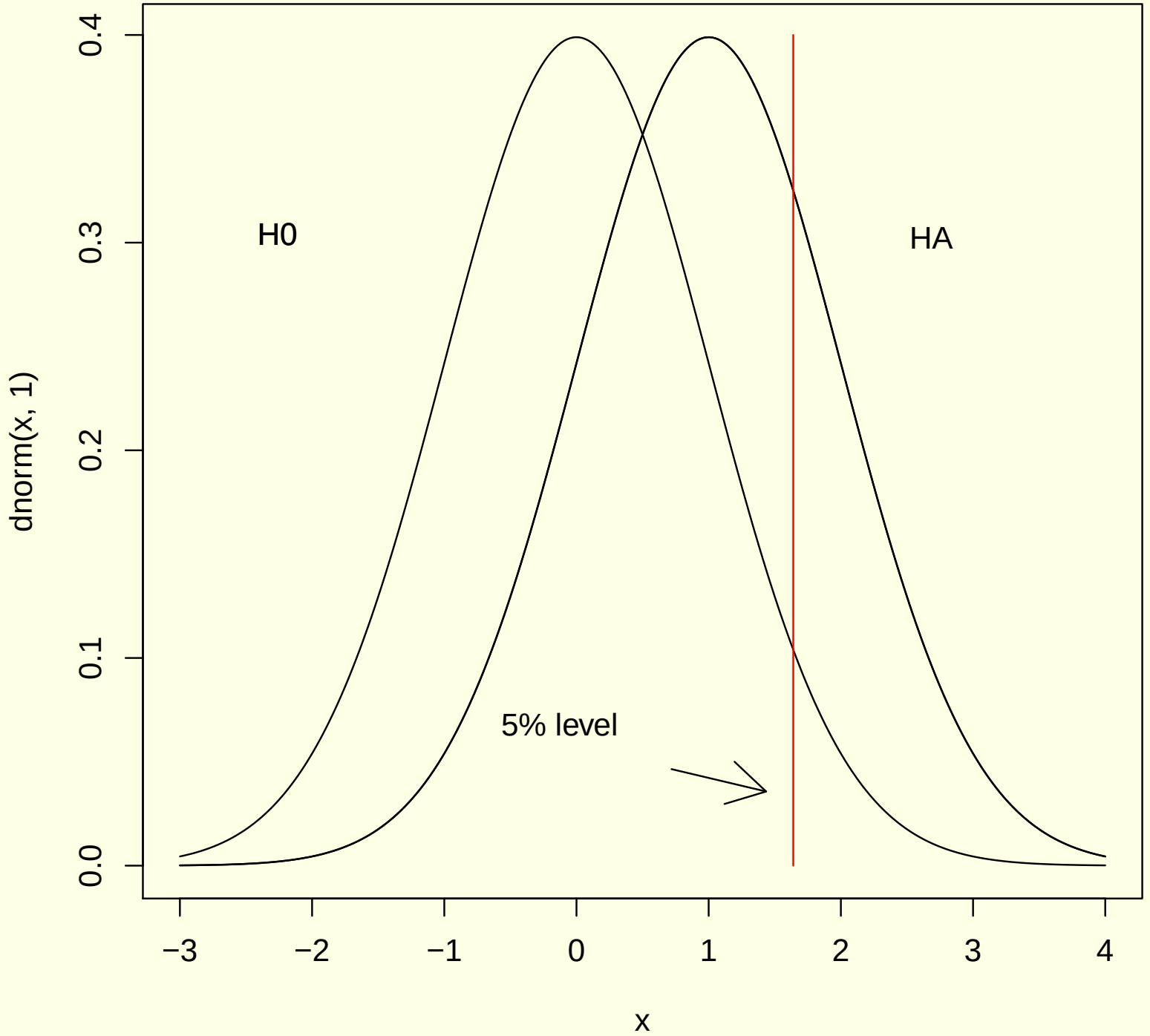
- But if the alternative were $X \sim N(0, .5)$, the test would be terrible: It would have higher probability of rejection under the null than under the alternative.

Pitfalls of the conventional 5% significance level

- Though there is increased awareness of its pitfalls among statisticians and social scientists, it is still very common for test results to be called "significant" when they reject at the 5% level and otherwise to be called "insignificant".

- Unless the data themselves have a two-point support (as they do in the disease-testing example), the results of a test by themselves usually suppress information in the likelihood function, even in univariate cases.

- Example: H0: $X \sim N(0,1)$, HA: $X \sim N(2,1)$. We construct a .05 level test by rejecting H when we observe $X > 1.64$. This results in an approximately 20% probabiiity under H0 that we accept H0 for data values that actually have higher probability under HA.


H0
HA
5% level
dnorm(x, 1)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
4
x

Inconsistency of tests with fixed significance level

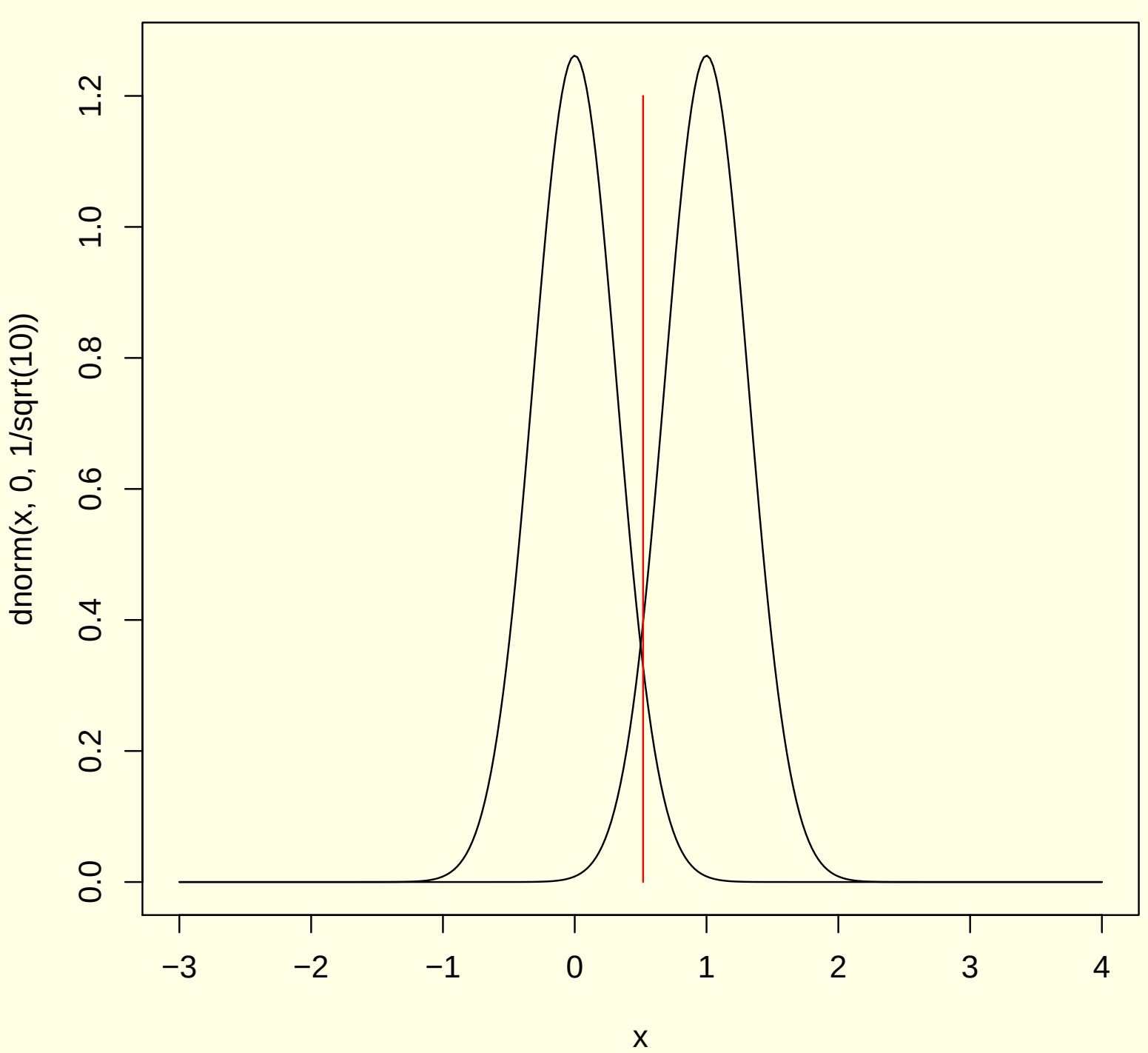
- If we test at the same significance level as sample sizes grow, the result is that we have a test procedure that is not **consistent.**

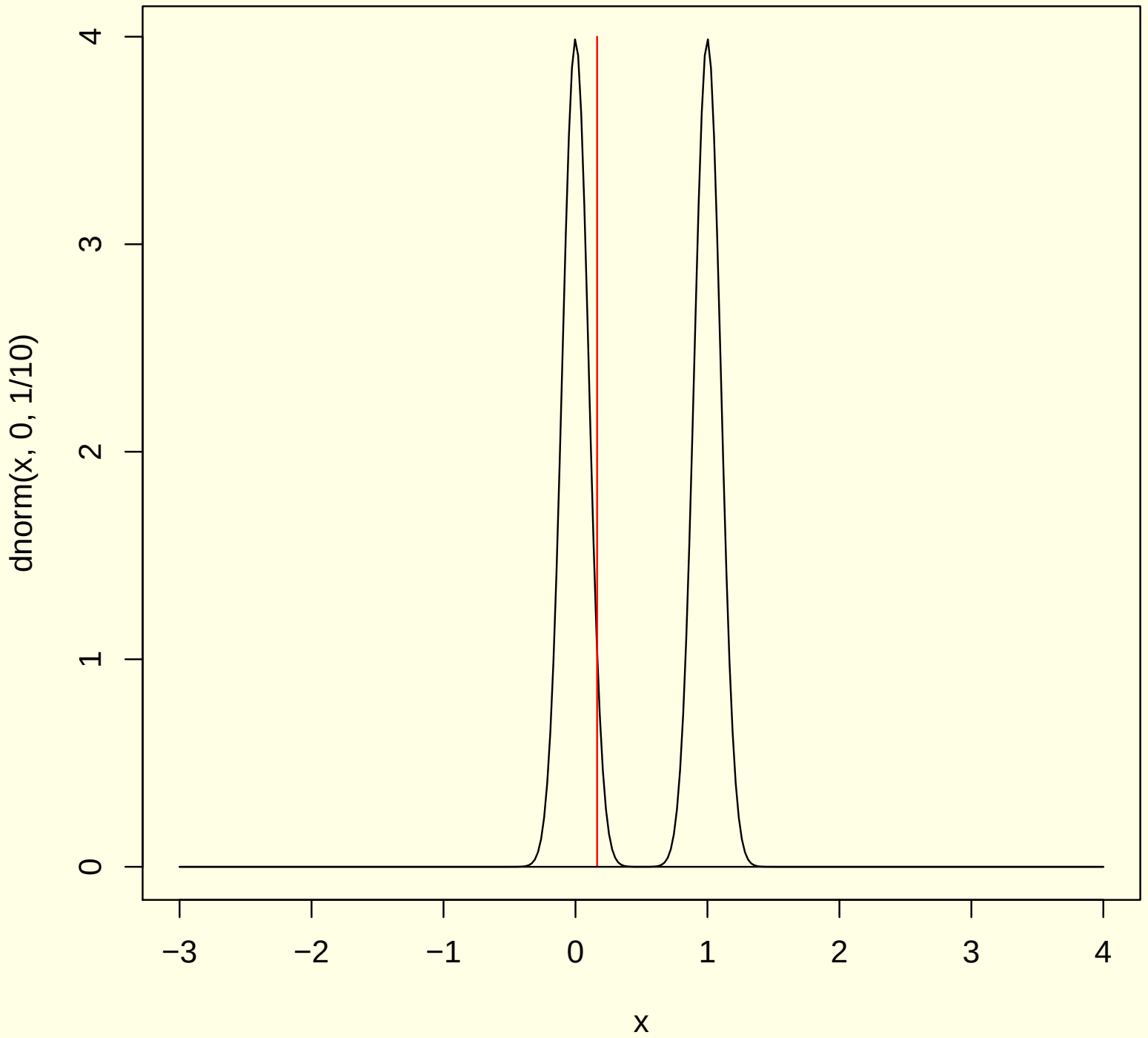
- A consistent test procedure is one that accepts a true null with probability approaching one as sample size grows and rejects a false null with probability approaching one as the sample size increases.

- Consistent test procedures generally require more stringent significance levels for larger samples.

Our example as N grows

In our example of a test for $N(0,1)$ vs. $N(1,1)$, in i.i.d. samples of size N, the posterior pdf would be centered on the mean, with standard deviation $1/\sqrt{N}$. The odds ratio favoring the null at the 5% level therefore goes to infinity with N. In large samples, much of the "rejection region" corresponds to odds ratios strongly favoring the null.


pdf's for xbar with sample size 10
dnorm(x, 0, 1/sqrt(10))
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−3
−2
−1
0
1
2
3
4
x


pdf's for xbar with sample size 100
dnorm(x, 0, 1/10)
0
1
2
3
4
−3
−2
−1
0
1
2
3
4
x

Compound H0 or HA

- For example: point null, compound alternative: H0: $X \sim N(0,1)$, HA: $X \sim N(\mu,1)$, $\mu \in \mathbb{R} - \{0\}$.
- Compound null, compound alternative: H0: $X \sim N(\mu,1)$, HA: $X \sim N(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+ - \{1\}$.
- The frequentist convention is to say that a test has significance level α when $\max_{\beta \in H0} P[\text{Regect} \mid \beta] = \alpha$.

Difficulties in interpreting compound tests

- Tests like this can be quite misleading, since their levels of type I and type II error can vary widely across parameter values in H0 and HA.
- The conventional definition of "significance level" is meant to be "conservative", in that the probability of rejecting the null is no greater than the significance level, thus favoring the null.
- But these "tests" are in fact used more often to characterize uncertainty about parameter values than to actually test some scientific hypothesis. Being "conservative" then can overstate uncertainty, which for decision making purposes can be as bad as understating it.

Credibility sets

- Suppose we have a model that provides a pdf $p(y \mid \beta)$ for every β in our parameter space S, and also a prior pdf $\pi(\beta)$.
- Then we can form a posterior pdf $q(\beta \mid Y)$ by Bayes rule.
- A set $C \subset S$ such that $P(C \mid Y) = 1 - \alpha$ is called a $100(1-\alpha)\%$ **credibility set** for β.
- The smallest such set has the form $\{\beta \mid q(\beta \mid Y) > \theta\}$, where θ is chosen so that the set has the desired probability. This is called an **HPD** credible set (for "highest posterior density").