

DSO 699: Statistics Theory

Special Topics in Data Sciences and Operations

Week 3
Adel Javanmard

USC Marshall
Department of Data Sciences and Operations

September 7, 2023

Last class

We talked about:

- **Global testing:** testing the global null

$$H_0 = \cap_{i=1}^n H_{0,i},$$

which holds if and only if all the individual nulls are true.

- The tale of two methods: Bonferroni's and Fisher's combination test
- When to use which?
- discussed the size of both methods
- Sharp detection boundary for the Bonferroni's method in the “needle in a haystack” model (sparse alternative)
- Optimality of Bonferroni's method?

Outline for today

- 1 χ^2 - test
- 2 Simes test
- 3 Tests based on empirical cdf

χ^2 - test

More on Fisher's combination test

Recall that for independent p -values p_i , the Fisher's test construct the statistics

$$T = -2 \sum_{i=1}^n \log p_i.$$

Our intuition is that this test is good for detecting ***many subtle changes***.

Can we formalize this intuition?

Consider our independent Gaussian sequence model

$$X_i = \mu_i + z_i, \quad z_i \stackrel{\text{i.i.d}}{\sim} N(0, 1)$$

$$H_0 : \mu_i = 0 \quad \forall i, \quad H_1 : \text{at least one } \mu_i \neq 0$$

χ^2 -test, a good approximation

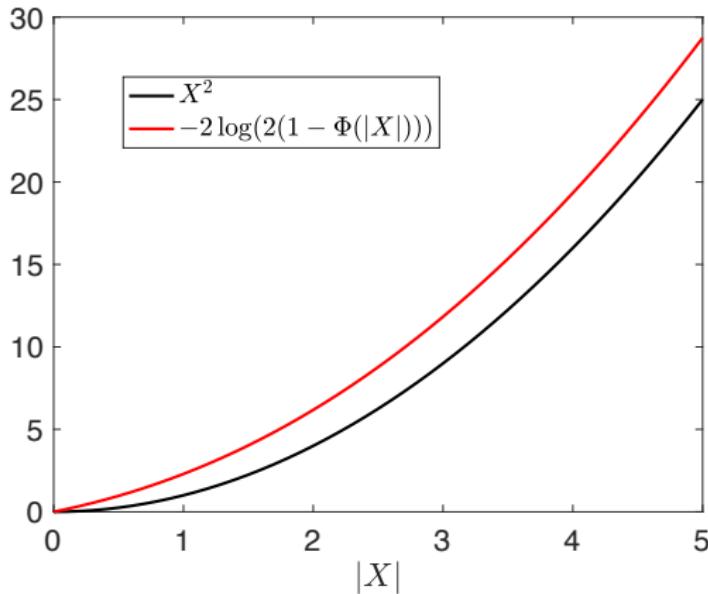
follows intuition.

nonzero mean

$$T^{\text{Fisher}} = -2 \sum_{i=1}^n \log p_i \quad T^{\text{chi}} = \sum_{i=1}^n X_i^2$$

-> nonzero data -> high Tchi value

same behavior (i think only under gaussian assumption)



Properties of χ^2 -test statistic

Under H_0 : $T_n \sim \chi_n^2$ and so the α -level test rejects H_0 when $T_n > \chi_n^2(1 - \alpha)$.

next page

Properties of χ^2 -test statistic

Under H_0 : $T_n \sim \chi_n^2$ and so the α -level test rejects H_0 when $T_n > \chi_n^2(1 - \alpha)$.

sum of iid squared standard normals

We can also use a Normal distribution approximation: By CLT, for large n we have

$$\frac{T_n - n}{\sqrt{2n}} \sim N(0, 1)$$

which in turn implies the following relation between quantiles of χ_n^2 and $N(0, 1)$:

$$\chi_n^2(1 - \alpha) \approx n + \sqrt{2n}z(1 - \alpha)$$

i.e. we can set critical value of p-value also with normal distn.
why we do that? to express critical value as simple function of n
and gaussian has simple expression for quantiles

Properties of χ^2 -test statistic

Under H_1 : T_n is non-central χ^2 :

$$T_n = \sum_{i=1}^n (\mu_i + z_i)^2$$

$$\mathbb{E}[(\mu_i + z_i)^2] = \mu_i^2 + 1$$

$$\text{Var}[(\mu_i + z_i)^2] = 4\mu_i^2 + 2$$

[next page](#)

Properties of χ^2 -test statistic

Under H_1 : T_n is non-central χ^2 :

$$T_n = \sum_{i=1}^n (\mu_i + z_i)^2 \quad \text{expand the square}$$

$$\mathbb{E}[(\mu_i + z_i)^2] = \mu_i^2 + 1$$

$$\text{Var}[(\mu_i + z_i)^2] = 4\mu_i^2 + 2$$

where one of μ_i is nonzero

Lyapunov CLT (not identical)

Applying CLT approximation, we have that for large n

$$\frac{T_n - (n + \|\mu\|^2)}{\sqrt{2n + 4\|\mu\|^2}} \sim N(0, 1)$$

subtract different mean and divide by different var
approximation of critical value and alternative distn are all based on gaussianity

Properties of χ^2 -test statistic

Rewriting the previous two slides:

Let $Z := \frac{T-n}{\sqrt{2n}}$ and

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \sqrt{\frac{n}{2}} \cdot \underbrace{\frac{\|\mu\|^2}{n}}_{\text{SNR}}$$

next page

Properties of χ^2 -test statistic

Rewriting the previous two slides: write as test statistic distribution;
not as data generating distribution

Let $Z := \frac{T-n}{\sqrt{2n}}$ and

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \sqrt{\frac{n}{2}} \cdot \underbrace{\frac{\|\mu\|^2}{n}}_{\text{SNR}} \quad \begin{array}{l} \text{signal to detect} \\ \text{noise} \\ \text{unit variance * n times} \end{array}$$

We then have (for large n)

$$H_0 : Z \sim N(0, 1) \quad \text{higher SNR means easy problem}$$

$$H_1 : Z \sim N\left(\theta, 1 + \frac{\theta}{\sqrt{n/8}}\right)$$

larger mean and larger variance

Power of χ^2 test

Power is given by

$$1 - \Phi \left(\frac{\Phi^{-1}(1 - \alpha) - \theta}{\sqrt{1 + \frac{\theta}{\sqrt{n/8}}}} \right)$$

(Why?)

next slide

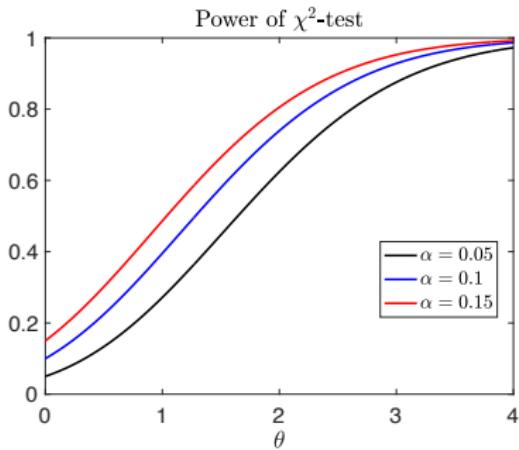
Power of χ^2 test

Power is given by

next slide

$$1 - \Phi \left(\frac{\Phi^{-1}(1 - \alpha) - \theta}{\sqrt{1 + \frac{\theta}{\sqrt{n/8}}}} \right)$$

(Why?)



Power of χ^2 test

Power is given by

goes alpha when
 $\theta > 0$

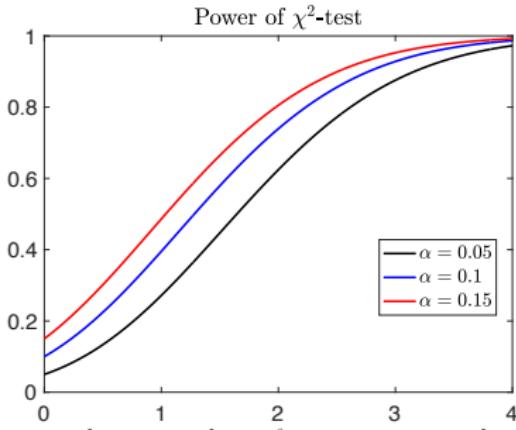
$$1 - \Phi \left(\frac{\Phi^{-1}(1 - \alpha) - \theta}{\sqrt{1 + \frac{\theta}{\sqrt{n/8}}}} \right)$$

different problem
different snr

(Why?)

so we can see
 θ is important
and θ is
l2 norm of means
divided by noise
(n_{test})

to make alt distn as standard normal



larger "size"
also increase the power

of course theta is unknown but we just want to see how power behaves wrt alt

- No sharp detection boundary!
- test is easy when $\theta \gg 1$ and hard when $\theta \ll 1$.

Optimality of the χ^2 test

Question: For the Gaussian sequence model, when $\theta \ll 1$, is there a test that does better than χ^2 ?

Optimality of the χ^2 test

Question: For the Gaussian sequence model, when $\theta \ll 1$, is there a test that does better than χ^2 ?

Answer: No! One can show that if $\theta \rightarrow 0$ as $n \rightarrow \infty$, then *for any test*

$$\liminf_{n \rightarrow \infty} \left(\mathbb{P}_{H_0}(\text{Type I error}) + \sup_{H_1} \mathbb{P}_{H_1}(\text{Type II error}) \right) \geq 1.$$

next slide

Optimality of the χ^2 test

Question: For the Gaussian sequence model, when $\theta \ll 1$, is there a test that does better than χ^2 ?

Answer: No! One can show that if $\theta \rightarrow 0$ as $n \rightarrow \infty$, then *for any test*

$$\liminf_{n \rightarrow \infty} \left(\mathbb{P}_{H_0}(\text{Type I error}) + \sup_{H_1} \mathbb{P}_{H_1}(\text{Type II error}) \right) \geq 1.$$

How?

- Same strategy we used to show optimality of Bonferroni's method.
- Use Neyman-Pearson Lemma
- Prove that LRT is powerless when $\theta \rightarrow 0$.
- More in HW 1

Comparison between Bonferroni's and χ^2 tests

Example 1. Suppose that $n^{1/4}$ of the μ_i 's are equal to $\sqrt{2 \log n}$.
(e.g., when $n = 10^6$, $n^{1/4} \approx 32$ and $\sqrt{2 \log n} \approx 5.3$)

In this case,

next slide

Comparison between Bonferroni's and χ^2 tests

bonferroni: threshold is set under the null,

so magnitude of $\max Y_i \sim \sqrt{\log n}$.

but when $n^{1/4}$ are this magnitude, prop of max being higher than thres is 1?

Example 1. Suppose that $n^{1/4}$ of the μ_i 's are equal to $\sqrt{2 \log n}$.

(e.g., when $n = 10^6$, $n^{1/4} \approx 32$ and $\sqrt{2 \log n} \approx 5.3$)

In this case,

NOT "only works well with few strong signal"

Rather, "works well with few or more strong signal"

- Bonferroni's test full power!
- χ^2 test has no power because

snr is

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \frac{n^{1/4} \times 2 \log n}{\sqrt{2n}} \rightarrow 0.$$

bonferoni sharp threshold does not apply here,
it is for one signal setting,

Comparison between Bonferroni's and χ^2 tests

Example 2. Suppose that $\sqrt{2n}$ of the μ_i 's are equal to 2 and the remaining ones are 0.

In this case,

[next slide](#)

Comparison between Bonferroni's and χ^2 tests

Example 2. Suppose that $\sqrt{2n}$ of the μ_i 's are equal to 2 and the remaining ones are 0.

In this case,

- χ^2 test has (almost) full power because

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \frac{4\sqrt{2n}}{\sqrt{2n}} = 4.$$

next slide

Comparison between Bonferroni's and χ^2 tests

Example 2. Suppose that $\sqrt{2n}$ of the μ_i 's are equal to 2 and the remaining ones are 0.

In this case,

- χ^2 test has (almost) full power because

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \frac{4\sqrt{2n}}{\sqrt{2n}} = 4.$$

- Bonferroni's test has no power. (Why?)

next slide

Comparison between Bonferroni's and χ^2 tests

Example 2. Suppose that $\sqrt{2n}$ of the μ_i 's are equal to 2 and the remaining ones are 0.

In this case,

- χ^2 test has (almost) full power because theta>1 easy. but precisely, look at the plot or the power

$$\theta = \frac{\|\mu\|^2}{\sqrt{2n}} = \frac{4\sqrt{2n}}{\sqrt{2n}} = 4.$$

remember that bf is comparing $\max x_i$ with $\sqrt{\log 2n}$ under OUR glob null

- Bonferroni's test has no power. (Why?)

✓ among the nulls, the largest X_i has size $\approx \sqrt{2 \log n}$

✓ among the non-nulls, the largest X_i has size $\approx 2 + \sqrt{2 \log \sqrt{2n}}$

✓ So the smallest p -values come from a null.

i.e. largest data will com from null, its because of the number
 $\max(A) = \max(\max(A') \max(A''))$ when n is large,
it's $\log n$ vs $\log(\sqrt{n})$
and the former wins

so two factors: larger mean and multiplicity

Simes test

personally, adel uses simes than bonferroni

only good thing of bf is it works on dependent p values

Simes test: another method for testing global null

As before, consider n hypotheses $H_{0,i}$ and p -values p_i .

We are interested in testing the global null $H_0 = \cap_{i=1}^n H_{0,i}$, where under $H_{0,i}$, $p_i \sim U[0, 1]$.

The Simes statistic is given by

$$T_n = \min_{1 \leq i \leq n} \left\{ p_{(i)} \frac{n}{i} \right\},$$

with ordered p -values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$.

It rejects the global null if $T_n \leq \alpha$.

inflate small p -values and deflate large p -value

Simes test: schematic illustration

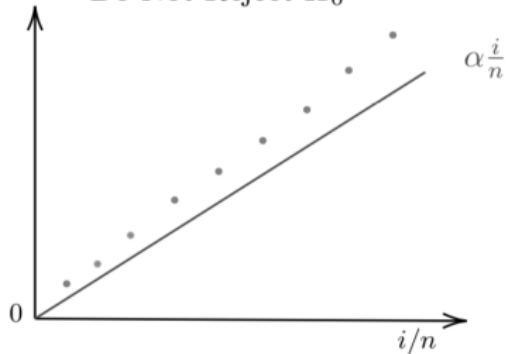
$$T_n := \min_{1 \leq i \leq n} \left\{ p_{(i)} \frac{n}{i} \right\} \leq \alpha \iff \exists i : p_{(i)} \leq \alpha \frac{i}{n}.$$

bonferroni

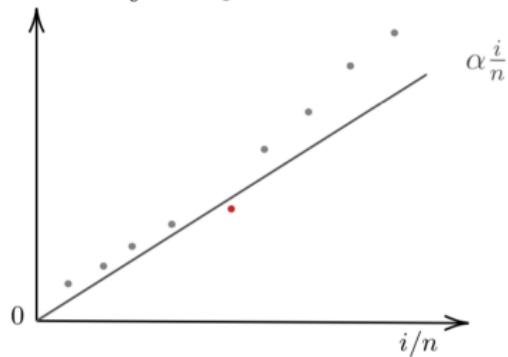
$\min p_i < a/n$

$\exists i : p_i < a/n < a/n$

Do Not Reject H_0



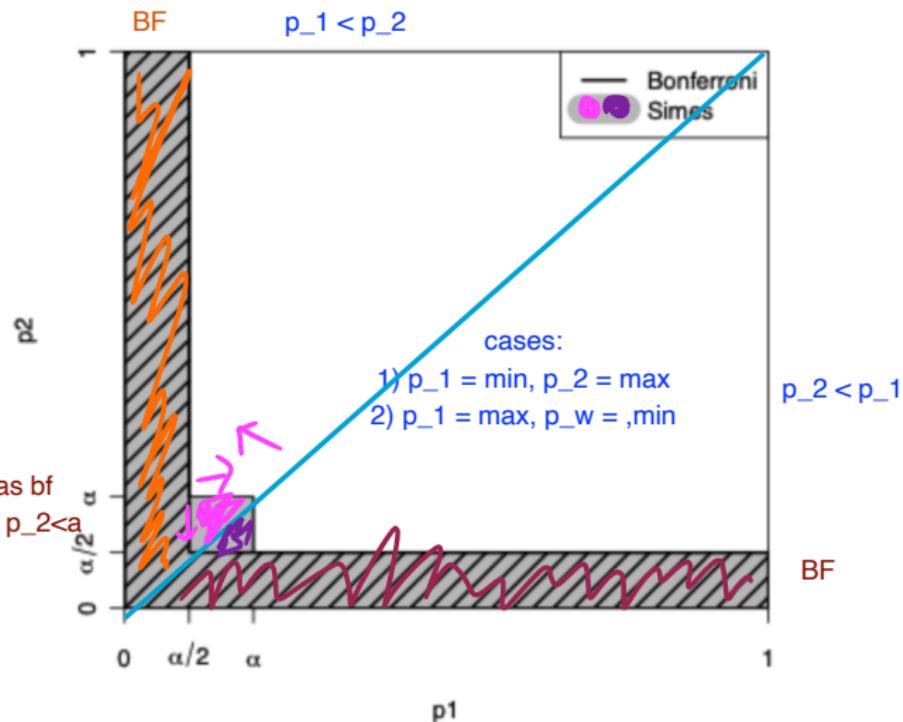
Reject H_0



from $i=2$, rejects more than bf
reject by bf implies reject by simes
while type I error is controlled (under independence)

Simes test versus Bonferroni's test

The Simes procedure is strictly less conservative than Bonferroni (Why?)



What is the area of the shaded region? $\alpha/2 + \alpha/2 - \alpha^2/4 + \alpha^2/4 = \alpha$.

if null and independent (points are uniform in the plane), Simes controls by a

Size of the Simes test

Theorem

If the individual p -values are independent, then under H_0 , $T_n \sim U[0, 1]$.

Thus, the size of Simes test is controlled at level α .

- Firstly note that $\mathbb{P}_{H_0}(p_{(n)} \leq t) = \mathbb{P}_{H_0}(\forall 1 \leq i \leq n, p_i \leq t) = t^n$.
So $p_{(n)}$ has density nt^{n-1} .
cdf of maximum p-value
cdf of maximum uniform, undergrad mathstat
- Secondly, we can write

$$\begin{aligned}\mathbb{P}(T_n \leq \alpha) &= \mathbb{P}\left(\min_{1 \leq i \leq n} \left\{ p_{(i)} \frac{n}{i} \right\} \leq \alpha\right) \\ &\quad \text{in first term,} \\ &= \mathbb{P}(p_{(n)} \leq \alpha) + \mathbb{P}(T_n \leq \alpha, p_{(n)} > \alpha) \\ &= \alpha^n + \mathbb{P}(p_{(n)} > \alpha, \min_{1 \leq i \leq n-1} \left\{ p_{(i)} \frac{n}{i} \right\} \leq \alpha)\end{aligned}$$

This sets the stage for induction on n .

Size of the Simes test (proof cont'd)

$n=1$ is same as bf, so base case
for induction

Conditioned on $p_{(n)} = t$, the other p -values are i.i.d from $U[0, t]$.
(so $p_i/t \sim U[0, 1]$) In addition, non-ordered

$$\min_{1 \leq i \leq n-1} \left\{ p_{(i)} \frac{n}{i} \right\} \leq \alpha \iff \min_{1 \leq i \leq n-1} \left\{ \frac{p_{(i)}}{t} \frac{n-1}{i} \right\} \leq \frac{\alpha}{t} \frac{n-1}{n}$$

ordered p-value are no longer uniform
correction of n into n-1

By applying the induction hypothesis,

interpret as new significance level

$$\begin{aligned} & \mathbb{P}(T_n \leq \alpha, p_{(n)} > \alpha) \\ &= \int_{\alpha}^1 \mathbb{P}(T_n \leq \alpha | p_{(n)} = t) nt^{n-1} dt \\ &= \int_{\alpha}^1 \mathbb{P}\left(\min_{1 \leq i \leq n-1} \left\{ p_{(i)} \frac{n}{i} \right\} \leq \alpha | p_{(n)} = t\right) nt^{n-1} dt \\ &= \int_{\alpha}^1 \mathbb{P}\left(\min_{1 \leq i \leq n-1} \left\{ \frac{p_{(i)}}{t} \frac{n-1}{i} \right\} \leq \frac{\alpha}{t} \frac{n-1}{n} | p_{(n)} = t\right) nt^{n-1} dt \\ &= \int_{\alpha}^1 \frac{\alpha}{t} \frac{n-1}{n} nt^{n-1} dt = \alpha - \alpha^n. \end{aligned}$$

suppose control in n-1 th step

Tests based on empirical cdf

Intuition

assumes independence

Recall the definition of empirical cdf of p_1, \dots, p_n given by

$$\hat{F}_n(t) = \frac{1}{n} \#\{i : p_i \leq t\}.$$

Under the global null $F(t) := \mathbb{P}(p_i \leq t) = t$, for $t \in [0, 1]$.

We would reject the global null hypothesis if the difference between $\hat{F}_n(t)$ and $F(t)$ is large as it is evidence against the null hypothesis.

We consider three tests based on the empirical cdf:

- The Kolmogorov-Smirnov Test
- Anderson-Darling Test
- Tukey's Second-Level Significance Test

Kolmogorov- Smirnov Test

The Kolmogorov- Smirnov (KS) test statistic is defined as

$$KS = \sup_{t \in \mathbb{R}} (\hat{F}_n(t) - F(t))$$

Note that small p -values (non-nulls) yields large values of $\hat{F}_n(t) - F(t)$.

Dvoretzky–Kiefer–Wolfowitz (DKW) inequality

Suppose $X_1, X_2 \dots, X_n \sim_{i.i.d} F(x)$. Then,

This holds for any F. Concentration inequality is meaningful only when it works with unknown distribution

$$\mathbb{P} \left(\sup_{x \in \mathbb{R}} (\hat{F}_n(x) - F(x)) > \varepsilon \right) \leq e^{-2n\varepsilon^2}, \quad \text{for every } \varepsilon \geq \sqrt{\frac{\ln 2}{2n}}.$$

this inequality handles sup over infinite without price, so useful for our setting
=a and solve

Therefore, to control the size of test at level α , we reject H_0 if

$$KS := \sup_{\substack{t \in \mathbb{R} \\ t \text{ in } [0,1], n \text{ values}}} (\hat{F}_n(t) - t) > \sqrt{\frac{1}{2n} \ln(1/\alpha)}$$

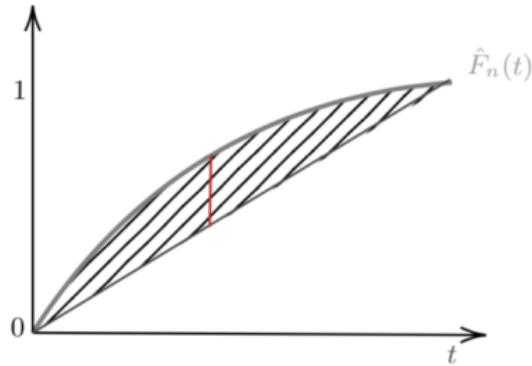
Anderson-Darling Test

Consider the test-statistic defined by

$$A^2 = n \int_0^1 (\hat{F}_n(t) - t)^2 \omega(t) dt,$$

where $\omega(t)$ is a weight function.

(if $\omega(t) = 1$ the above statistic is called the Cramer-von Mises statistic)



Anderson-Darling Test (cont'd)

Anderson-Darling chooses the weight function $\omega(t) = [t(1 - t)]^{-1}$.

$$A^2 = n \int_0^1 \underbrace{\frac{(\hat{F}_n(t) - t)^2}{t(1 - t)}}_{\text{squared z-score}} dt$$

Puts more weight on small/ large p -values.

(There are specific nasty formula to calculate p -values based on A^2)

[Jantschi, Bolboacă, 2018]

Higher-criticism

(a.k.a Tukey's Second-Level Significance Testing)

"A young psychologist administers many hypothesis tests as part of a research project, and finds that, of 250 tests 11 were significant at the 5% level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher (Tukey himself?) suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance."

[D. Donoho and J. Jin 2004]

Higher-criticism

(a.k.a Tukey's Second-Level Significance Testing)

"A young psychologist administers many hypothesis tests as part of a research project, and finds that, of 250 tests 11 were significant at the 5% level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher (Tukey himself?) suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance."

[D. Donoho and J. Jin 2004]

think as biased coin

He then proposed the statistic

$$HC_{0.05,n} = \frac{n\widehat{F}_n(0.05) - 0.05n}{\sqrt{n0.05(1 - 0.05)}}$$

beroulli sd

and suggested that values of (say > 2) indicate a *significance of the overall body of tests!*

if under null, p-values are uniform, so rejecting the null is $Ber(0.05)$

nF_n is observed data of number of rejection

Higher-criticism (cont'd)

The high-criticism constructs the statistic:

$$\max_{0 \leq \alpha \leq \alpha_0} \frac{\sqrt{n}(\hat{F}_n(\alpha) - \alpha)}{\sqrt{\alpha(1 - \alpha)}},$$

for some $\alpha_0 > 0$.

(For Comparison, recall Anderson-Darling statistic:)

$$A^2 = n \int_0^1 \frac{(\hat{F}_n(t) - t)^2}{t(1 - t)} dt$$

L_{∞} vs squared L_2

Recap

We talked about

- χ^2 -test
- Power and size of χ^2 -test (no sharp detection boundary)
- Simes test (and proof of its validity)
- Tests based on empirical cdf of p -values and how much it diverges from a uniform distribution:
 - ✓ The Kolmogorov-Smirnov Test
 - ✓ Anderson-Darling Test
 - ✓ Tukey's Higher-Criticism Test

