TODAY

1. Null Hypothesis Testing / p-values
2. Simple Refined MDL with Simple H_0 as Null Hypothesis Testing
   - MDL provides **always-valid** p-values
3. Financial Interpretation of MDL with Simple H_0 - Kelly Gambling
4. What about Composite H_0?

[Most of this not in book – these slides are reference material]

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, today we assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}, \Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0\}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1\}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}, \Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$

  Simple H_0

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  H_1 : $X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown (‘nuisance’) parameter

  $$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$$
  $$H_1 = \left\{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \right\}$$

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  Composite H_0

  $H_1: X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$$
  $$H_1 = \{ P_{\sigma,\mu} \big| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$$

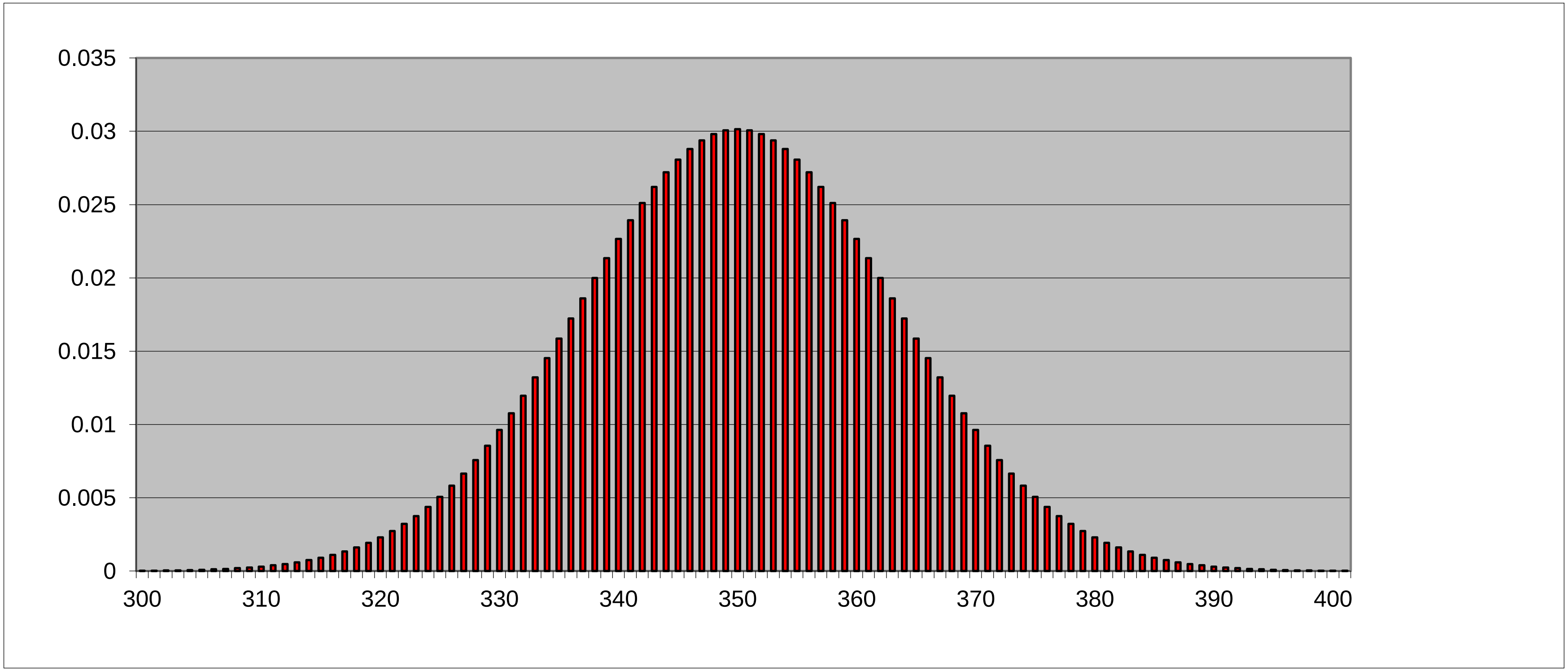
Standard Method: p-value, significance

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
- A ("nonstrict") **p**-value is a random **variable** (!) such that, for all $\theta \in \Theta_0$,

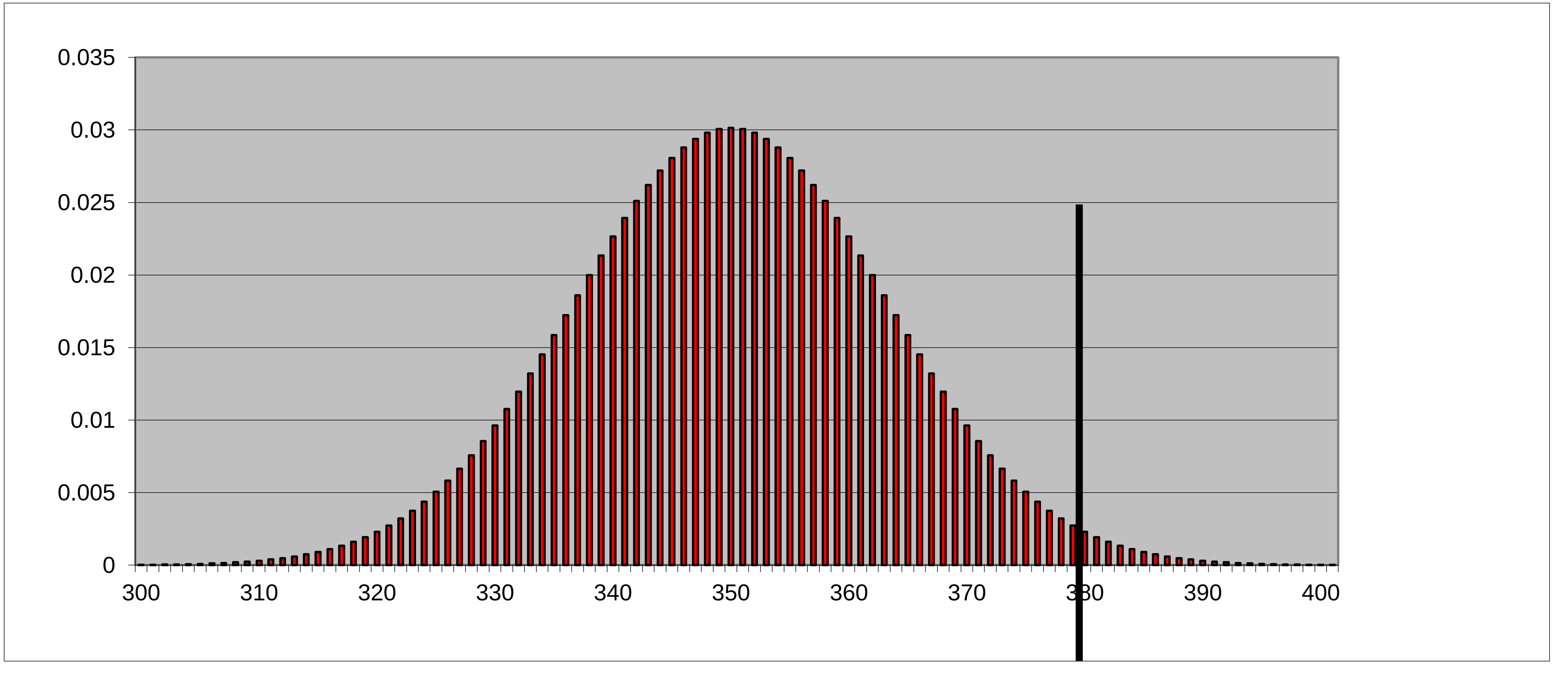
$$P_{\theta_0}(\mathbf{p} \leq \alpha) \leq \alpha$$

Coin Tossing Example, $n = 700$

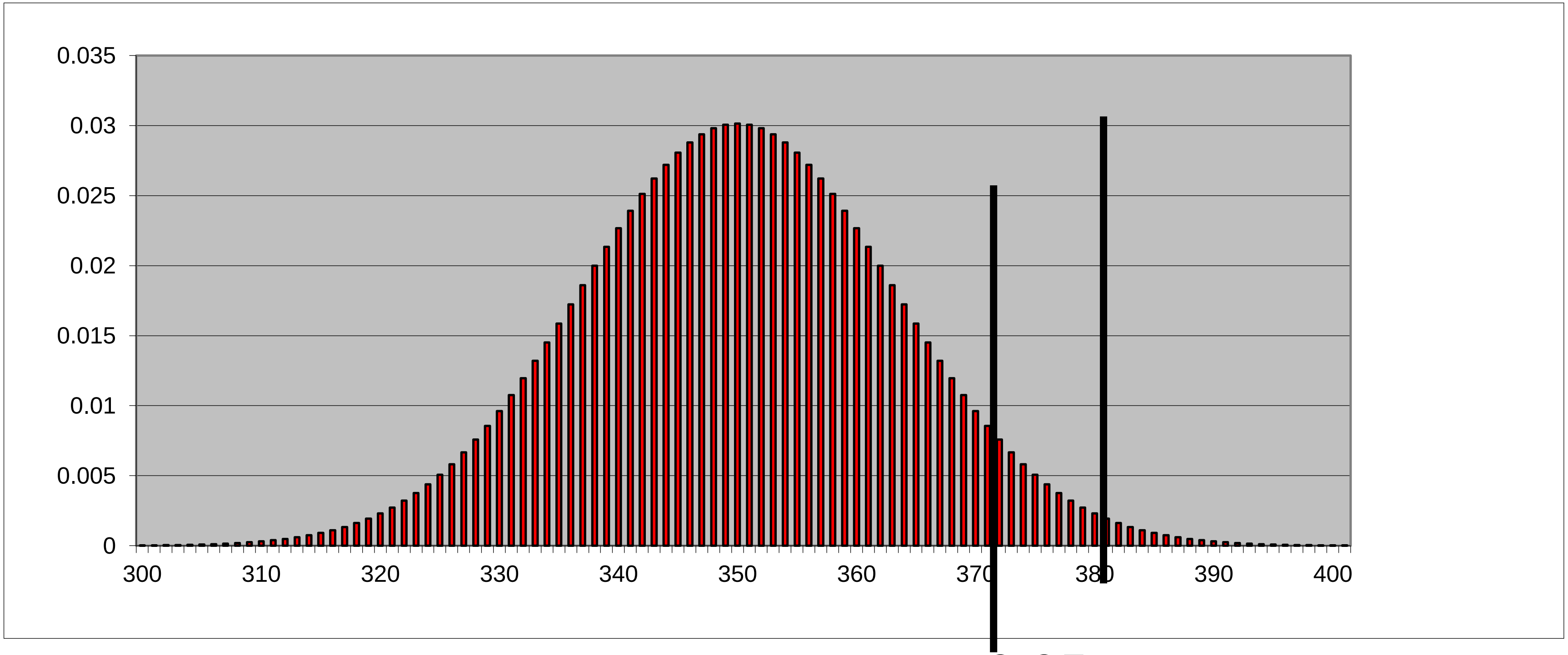
According to H_0 : $T := \sum_{i=1}^{700} X_i \sim \text{Bin}(0.5, 700)$



- We now do an experiment and we observe *T=380*. **The p-value is the probability that we would get this value, or an even smaller one**
- ≈ total probability mass right from black line. We find, for *T = 380,* that *p = 0.02*



- We determine (before experiment!) a ***significance level*** α and we 'reject' the null hypothesis iff $p \leq \alpha$
- This gives a **Type-I Error Probability bound** α
- **If we follow this decision rule consistently throughout our lives, then in long run we reject the null while it is correct at most 5% of the time**



p = 0:05

Significance Testing

- The Significance Test against H_0 at level α based on p-value p is defined as the test which rejects H_0 if $p = p(X^n) \leq \alpha$
- Thus a level α – test has **Type-I Error** Bound of 0.05

$$P\left(\text{“Test says reject”}\right) = P\left(p \leq \alpha\right) \leq \alpha$$

Simple Refined MDL and Hypothesis Testing

Given $H_0 = \{ p_\theta | \theta \in \Theta_0 \}$ vs $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:
Evidence in favour of H_1 measured by

$$\bar{L}_0(X^n) - \bar{L}_1(X^n) = \log \frac{\bar{p}_1(X_1, \ldots, X_n)}{\bar{p}_0(X_1, \ldots, X_n)}$$

where $\bar{p}_j(X_1, \ldots, X_n)$ represents universal distribution relative to H_j e.g.

$$\bar{p}_j(X_1, \ldots, X_n) = \int_{\theta \in \Theta_j} p_\theta(X_1, \ldots, X_n) w_j(\theta) d\theta$$

or

$$\bar{p}_j(X_1, \ldots, X_n) = \bar{p}_{\mathsf{nml},j}(X^n) = \frac{p_{\hat{\theta}_j(X^n)}(X^n)}{\sum_{z^n \in \mathcal{X}^n} p_{\hat{\theta}_j(z^n)}(z^n)}$$

Simple Refined MDL, simple $\boldsymbol{H_0}$

MDL hypothesis testing

between $H_0 = \{ p_0\}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1\}$:

Evidence in favor of H_1 measured by

$\log M(X^n)$ where $M(X^n) := \dfrac{\bar{p}_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$

...since the only reasonable 'universal' distribution relative to $\boldsymbol{H_0}$ is $\boldsymbol{p_0}$ itself

Simple Refined MDL, simple H_0

Evidence in favor of H_1 measured by

$$\log M(X^n) \text{ where } \quad M(X^n) := \frac{\bar{p}_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

Note that

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] = \int p_0(x^n) \cdot \frac{\bar{p}_1(X^n)}{p_0(x^n)} dx^n = \int \bar{p}_1(x^n) dx^n = 1$$

Hence by Markov's Inequality

$$P_0(M^{-1}(X^n) \leq \alpha) = P_0(M(X^n) \geq \alpha^{-1})$$

$$\leq \frac{\mathbf{E}_{X^n \sim P_0}[M(X^n)]}{\alpha^{-1}} \leq \alpha$$

Simple Refined MDL, simple H_0

...so, no matter how $\bar{p}_1$ is defined,

(1) The MDL Evidence for Simple H_0 provides a p-value!

(2) Thus (see next slide) Selecting H_1 (i.e. Rejecting H_0) if

$$\bar{L}_0(X^n) - \bar{L}_1(X^n) \geq \log 20$$

gives a classical null hypothesis test with significance level $\frac{1}{20} = 0.05$

$$\leq \frac{\mathbf{E}[M(X^n)]}{\alpha^{-1}} \leq \alpha$$

We have just seen that, no matter how $\bar{p}_1$ is defined :

$$P_0\left(-\log p_0(X^n) - [-\log \bar{p}_1(X^n)] \geq -\log \alpha\right) =$$

$$P_0\left(\frac{\bar{p}_1(X^n)}{p_0(X^n)} \geq \alpha^{-1}\right) \leq \frac{\mathbf{E}_{P_0}\left[\frac{\bar{p}_1(X^n)}{p_0(X^n)}\right]}{\alpha^{-1}} = \alpha.$$

Simple Refined MDL, simple H_0

...so, no matter how $\bar{p}_1$ is defined,

(1) The MDL Evidence for Simple H_0 provides a p-value!

(2) Thus selecting H_1 (i.e. Rejecting H_0) if

$$\bar{L}_0(X^n) - \bar{L}_1(X^n) \geq \log 20$$

gives a classical null hypothesis test with significance level $\frac{1}{20} = 0.05$

$$\leq \frac{\mathbf{E}[M(X^n)]}{\alpha^{-1}} \leq \alpha$$

The fact that MDL with simple H_0 provides a p-value is just the No Hyper-Compression Inequality

- We have just seen that, no matter how $\bar{p}_1$ is defined :

$$P_0\left(-\log p_0(X^n) - [-\log \bar{p}_1(X^n)] \geq -\log \alpha\right) \leq \alpha$$

- i.e. (set $\alpha = 2^{-K}$, $-\log \alpha = K$) the probability that with any code $\bar{L}_1$ we can compress data coming from $\bar{p}_0$ by K bits or more compared to the best code for $\bar{p}_0$ is bounded by 2^{-K}

- This is just a generalization of no-hypercompression inequality: we saw and proved this for P_0 is Bernoulli (1/2) in the very first lecture

Better No-Hypercompression (not in book)

$$P_0\left(\exists n : \frac{p_0(X^n)}{\bar{p}_1(X^n)} \leq \alpha\right) = P_0(\exists n : M(X^n) \geq \alpha^{-1}) \leq \alpha$$

- Proof: Special Case of Doob's Optional Stopping Theorem (1949)
- Intuitive Reason:
  - (1) the exponentiated codelength difference (i.e. likelihood ratio) can be interpreted in terms of money (Kelly gambling)
  - (2) No matter what your rule is for when to go home, you don't expect to make money in a casino

Data Compression as
Gambling!

Data Compression as Gambling!

Kelly (1956)

- At time 1 you can buy ticket 1 for 1$. It pays off $M_1 = \bar{p}_1(X_1)/p_0(X_1)$$
- At time 2 you can buy ticket 2 for 1$. It pays off $M_2 = \bar{p}_1(X_2|X^1)/p_0(X_2)$ $ .... and so on.

  You may buy multiple and fractional nrs of tickets.

- You start by investing 1$ in ticket 1.
- After 1 outcome you either stop with end capital M_1 or you continue and buy M_1 tickets for round 2. After second round you stop with end capital $M_1 \cdot M_2$ or you continue and buy $M_1 \cdot M_2$ tickets for third round, and so on..

- You start by investing 1$ in ticket 1.
- After 1 outcome you either stop with end capital M_1 or you continue and buy M_1 tickets for round 2. After second round you stop with end capital $M_1 \cdot M_2$ or you continue and buy $M_1 \cdot M_2$ tickets for third round, and so on..
- M_n is simply your accumulated capital after n rounds
- If null hypothesis true, then at each round, you do not expect to increase your wealth:

$$\mathbf{E}_{P_0}[M_n \mid X^{n-1}] = \mathbf{E}_{P_0}\left[\frac{\bar{p}_1(X_n|X^{n-1})}{p_0(X_n)}\right] = 1$$

- M_n is simply your accumulated capital after n rounds
- If null hypothesis true, then at each round, you do not expect to increase your wealth:

$$\mathbf{E}_{P_0}[M_n \mid X^{n-1}] = \mathbf{E}_{P_0}\left[\frac{\bar{p}_1(X_n|X^{n-1})}{p_0(X_n)}\right] = 1$$

- ...so the fact that “the probability that you ***ever*** gain more than \$20 is bounded by 1/20 “ is simply a formalization of the common knowledge that ‘it’s unlikely that you get rich in a casino, no matter what rule you use to decide when to go home!’

Data Compression as Gambling!

Kelly (1956)

- At time j you can buy ticket j for 1\$. It pays off $M_j = \bar{p}_1(X_j|X^{j-1})/p_0(X_j)$\$
- Equivalent, more intuitive view: let $\mathcal{X} = \{1, \ldots, K\}$. At time j there are K tickets available. Ticket k pays off $1/p_0(k)$ if outcome is k, and 0 otherwise.
- You think of $\bar{p}_1(\cdot|X^{j-1})$ as a **strategy** for dividing your capital over the K tickets: you put a fraction $\bar{p}_1(X_j = k|X^{j-1})$ of your money obtained so far on ticket K
- Then your total capital gets multiplied by $M_j = \bar{p}_1(X_j|X^{j-1})/p_0(X_j)$

Data Compression as Gambling!

Kelly (1956)

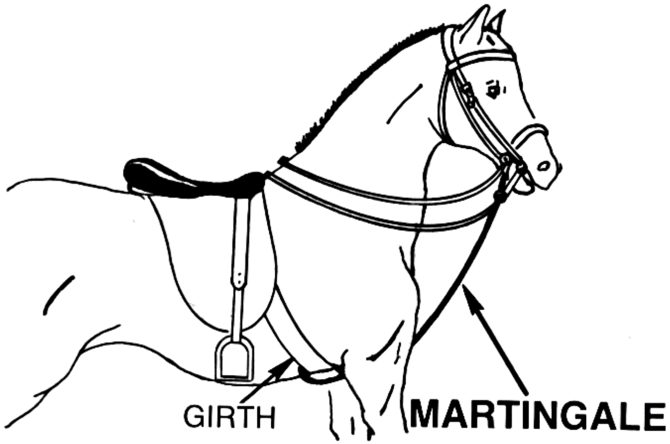
- Standard interpretation $-\log \bar{p}_1$ and $-\log p_0$ are both code-lengths.
- New interpretation:

  $\bar{p}_1$ is investment strategy, p_0 determines pay-offs

  [or vice-versa!]

Technical Aside (for those who know stochastic process theory)

- Technically, we can view the process $(M_1, M_1 \cdot M_2, M_1 \cdot M_2 \cdot M_3, \ldots)$ as a nonnegative supermartingale.
- The Type-I Error Probability result is then Ville's (1939) Inequality, and the Proof is Immediate by Doob's Optional Stopping Theorem



MDL Model Selection with Simple Null

- Codelength difference, or equivalently, likelihood ratio, also gives ‘robust’ (**always-valid**) p-value
- Less sharp than standard p-value
  - you need more data to get significant result
  - …but you get something back for that: you can stop/continue whenever you want
- …the fact that you cannot do optional stopping with p-value is **one of the major reasons for the replicability crisis in science**

- But what about **composite null?**

Composite H_0: Simple Refined MDL does not always give an always-valid p-value

Say, $\bar{p}_0 = \bar{p}_{W_0}$ is Bayesian universal distribution.
Evidence given by

$$M(X^n) := \frac{\bar{p}_1(X_1, \ldots, X_n)}{\bar{p}_{W_0}(X_1, \ldots, X_n)}$$

No Hypercompression/p-value interpretation requires that **for all** $P_0 \in H_0$:

$$\mathbf{E}_{X^n \sim P_0} [M(X^n)] \leq 1$$

...but we can only guarantee "average statement" that

$$\mathbf{E}_{X^n \sim \bar{P}_{W_0}} [M(X^n)] = \mathbf{E}_{\theta \sim W_0} \mathbf{E}_{X^n \sim P_\theta} [M(X^n)] \leq 1$$

Composite H_0: Simple Refined MDL does not always give an always-valid p-value

- In general MDL with composite H_0 does not give p-values let alone always-valid p-values
- ...but there do exist *very special priors* W_1^* **,** W_2^* **(sometimes highly unlike priors that "Bayesian" statisticians tend to use!)** for which $\bar{p}_{W_1^*}, \bar{p}_{W_0^*}$ provide universal distributions such that the corresponding likelihood/MDL ratio does give an always-valid p-value

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

σ^2 unknown ('nuisance') parameter

$H_0 = \{ P_\sigma | \sigma \in (0, \infty)\}$ $\quad H_1 = \left\{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\}\right\}$

- In general Bayes factors are not S-values
- But lo and behold, Jeffreys' uses very special priors and his Bayes factor is an S-value, so his Bayesian t-test is a Safe Test!

Example:
Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

Jeffreys uses improper right-Haar prior $w(\sigma) = 1/\sigma$ within both models, and uses Cauchy on $\delta := \mu/\sigma$

$$\bar{p}_0(X^n) := \int_{\sigma>0} w(\sigma) p_{0,\sigma}(X^n) d\sigma = \int \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \frac{1}{\sigma} \cdot \exp\left(-\frac{\sum X_i^2}{2\sigma^2}\right) d\sigma$$

$$\bar{p}_1(X^n) := \int_{\delta\in\mathbb{R},\sigma>0} w(\delta) w(\sigma) p_{\delta,\sigma}(X^n) d\sigma d\delta, \;\; p_{\delta,\sigma}(X^n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i}{\sigma} - \delta\right)^2\right)$$

- With this choice $S := \bar{p}_1(X^n)/\bar{p}_0(X^n)$ has same distribution under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}[S] = 1$

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1: X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

- Jeffreys uses improper right-Haar prior $w(\sigma) = 1/\sigma$ within both models, and uses $w_{[\delta]}$ Cauchy on $\delta = \frac{\mu}{\sigma}$
- In fact, for right-Haar prior combined with arbitrary prior on effect size $\delta = \mu/\sigma$ we get that S has same distr. under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}(S) = 1$

Example 2: Independence Testing/2x2 tables

- $X_i \in \{0,1\}$; $Z_i \in \{m, f\}$

- H_0: $X_1, X_2, \dots, X_n \mid Z_1, \dots, Z_n$ iid Bernoulli(θ),
- H_1: $X_1, X_2, \dots, X_n \mid Z_1, \dots, Z_n$ independent but
  $P(\, X_i = 1 \mid Z_i = m \,) = \theta_m$
  $P(\, X_i = 1 \mid Z_i = f \,) = \theta_f \neq \theta_m$
- Are **both populations same or different?**

2x2 Contingency Tables

- For $\Theta_1 = \{ (\theta_f, \theta_m) \in [0,1]^2\}$

$$M_n := \frac{\int_\delta w(\delta) p_{(1/2+\delta,1/2-\delta)}(x^n \mid z^n) d\delta}{p_{1/2}(z^n)}$$

…gives an always-valid p-value, for every prior density on $\delta \in [-\frac{1}{2}, \frac{1}{2}]$

next week:

Safe Testing

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

with Rianne de Heide, Wouter Koolen, Judith ter Schure, Alexander Ly, Rosanne Turner




P-value Problem: Combining Dependent Tests

- Suppose reseach group A tests medication, gets ‘almost significant’ result.
- ...whence group B tries again on new data. How to combine their test results?
    - **Standard methods for combining p-values (Fisher’s and Stouffer’s) require independence hence cannot be applied**
- **With the type of “p-value” introduced here, despite dependence, evidences can still be safely multiplied**

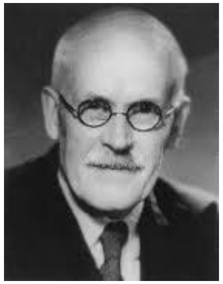
P-value Problem (b): Extending Your Test

- Suppose reseach group A tests medication, gets 'almost significant' result.
- **Sometimes group A can't resist to test a few more subjects themselves...**
  - A recent survey revealed that **55% of psychologists** have succumbed to this practice
- But isn't this just **cheating?**
  - **Not clear: what if you submit a paper and the *referee* asks you to test a couple more subjects? Should you refuse because it invalidates your p-values!?**

Three Philosophies of Testing

Jerzy Neyman: alternative exists, "inductive . . behaviour", 'significance level' and power

Sir Ronald Fisher: test statistic rather than alternative, p-value indicates "unlikeliness"



Sir Harold Jeffreys: **Bayesian**, alternative exists, absolutely no p-values

J. Berger (2003, IMS Medaillion Lecture): ***Could Neyman, Fisher and Jeffreys have agreed on testing?***

... Using always-valid p-values based on MDL we can unify/correct the central ideas