t-student

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\hat{sd}}$$

$\hat{sd}$ is the estimation of the standard deviation.

$$\hat{sd} = \hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad \text{... use ①}$$

↳ Using t-student, the estimator does not depend on the other parameters of the distribution

↳ it allows to test $\bar{x}_1 - \bar{x}_2$ (and come with confidence intervals) without the estimation of $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^{2?}$

t-test statistic is a pivotal statistic as it does not depend on other parameters the distribution might have.

★ So. Frequentist use pivotal statistics whenever they are available to conduct stat tests.

27 Aug 2019

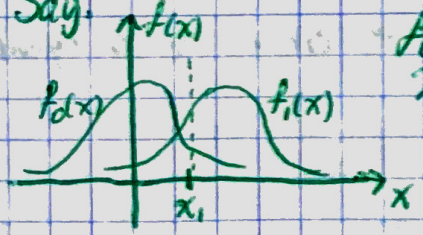
Frequentist Optimality

- Neyman-Pearson lemma provides an optimum hypothesis-testing algorithm.

→ NP lemma states:

1) if we are to decide between 2 possible density functions for observed data x
   → a null hypothesis density $f_0(x)$,
   → and an alternative density $f_1(x)$

Say.



for the simple case $x = x_1$

So, with what confidence level can we say x_1 does not belong to f_0 (regeacting H_0)?

Determine the hypothesis test that is powerful at:
regeacting, but not regeacting when it's the appropriate thing.
(no false positives)

ii) A testing rule $t(x)$ says which choice (0 or 1) we'll make given x.

For any decision we make, there will be two associated errors α and β

$\alpha = Pr_{f_0}\{t(x)=1\}$ ⇒ the probability of rejecting the null hypothesis when x belongs to the null hypothesis (incorrectly rejecting f_0)

$\beta = Pr_{f_1}\{t(x)=0\}$ ⇒ choosing f_0 when actually f_1 generated x probability of (incorrectly failing to reject)

iii) Let $L(x)$ be the Likelihood ratio

$$L(x) = \frac{f_1(x)}{f_0(x)}$$

what is the likelihood of the alternative hypothesis $f_1(x)$ giving the data x vs. what is the likelihood of the null hypothesis $f_0(x)$ giving your data x

iv) Testing rule $t_c(x)$ by:

$$t_c(x) = \begin{cases} 1 & \text{if } \ln(L(x)) \geq c' \\ 0 & \text{if } \ln(L(x)) < c' \end{cases}$$

As only rules of the $t_c(x)$ form can be optimal

there is one such rule for each choice of the cutoff c'.

→ that is $\alpha_c < \alpha$ & $\beta_c < \beta$ for some c and where α & β are generated by another rule.

if $t_c(x)$ has an α_c with a c' such that the α error α_c is equal to α, then $\beta_c < \beta$ (and vice versa)

e.g.

We want to get c such that α is equal to Type I error rate and Type II error β is minimized.

$f_0 \sim N(0,1)$ ⚡ $f_1 \sim CN(1/2, 1)$

$f_0(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$ ⚡ $f_1(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - 1/2)^2}{2}}$ ⚡ for normal dists density functions.

$f_0(x) = \left[\frac{1}{\sqrt{2\pi}}\right]^n \prod_{i=1}^{n} e^{\frac{-x_i^2}{2}} = \left[\frac{1}{\sqrt{2\pi}}\right]^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}$

$f_1(x) = \left[\frac{1}{\sqrt{2\pi}}\right]^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - 1/2)^2}$

} the likelihood functions

$L(x) = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - 1/2)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}}$ ⚡ likelihood ratio

$L(x) = e^{-\frac{1}{2}\left[\sum_{i=1}^{n} x_i + \frac{n}{4}\right]}$

$L(x) = e^{-\frac{1}{2}\left[n\bar{x} + \frac{n}{4}\right]}$

$L(x) > c \Rightarrow e^{-\frac{1}{2}\left[n\bar{x} + \frac{n}{4}\right]} > c \Rightarrow \bar{x}$ shall be greater than some constant.

$-\frac{1}{2}\left[n\bar{x} + \frac{n}{4}\right] > c_1$

$n\bar{x} + n/4 > c_2$

$\bar{x} > c_3$ ⟶ only the mean depends on the sample.

» If two likelihood functions come from a normal distribution, then the only thing we need for the hypothesis testing is the sample mean $\bar{x}$

» Most powerful hypothesis test at any Type I error rate α is to compare $\bar{x}$ to a constant.

→ let's relate c, α & β...

$\alpha = P(\bar{x} > c \mid \theta = 0)$ where $\theta = \mu = 0$

↳ probability under $f_0 \sim N(0, \sigma^2)$ where $\sigma^2 = 1$

↳ $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{1}{n} \Rightarrow \sigma_{\bar{x}}^2 = \frac{1}{\sqrt{n}}$

then $\frac{\bar{x} - 0}{1/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}^2} = \bar{x}\sqrt{n} \sim N(0,1)$

normalize

- $\alpha = P(\bar{x} > c \mid \mu = 0)$

$\alpha = P(\bar{x}\sqrt{n} > c\sqrt{n} \mid \mu = 0)$

$\alpha = 1 - P(\bar{x}\sqrt{n} \leq c\sqrt{n} \mid \mu = 0)$

$\alpha = 1 - \Phi(c\sqrt{n})$

↳ function for the CDF of a $N(0,1)$ distribution.

Φ is the cumulative density function (CDF) of a normal distribution.

$\Phi(c\sqrt{n}) = 1 - \alpha$

$c\sqrt{n} = \Phi^{-1}(1-\alpha)$

$c = \frac{1}{\sqrt{n}} \Phi^{-1}(1-\alpha)$

In general: $c = \mu_0 + \frac{1}{\sqrt{n}} \Phi^{-1}(1-\alpha)$ for two normal distributions with the same variance

- $1-\beta$: Power of hypothesis test

↳ Probability of correctly rejecting

- A hypothesis test for which $1-\beta$ is maximized (β is minimized) $\hat{\alpha}$ level is called a Most Powerful Test (MPT)

On the other hand:

- $\beta = P(\bar{x} \leq c \mid \mu = 1/2)$

↳ fail to reject the null hypothesis when the alternative hypothesis is true.

$\frac{\bar{x} - \frac{1}{2}}{\frac{1}{\sqrt{n}}} \sim N(0,1)$ } normalize to a $N(0,1)$ distribution

$\beta = P[(\bar{x} - 1/2) \cdot \sqrt{n} \leq (c - 1/2) \cdot \sqrt{n} \mid \mu = 1/2]$

$\beta = \Phi((c - 1/2)\sqrt{n})$

Smallest Type II error rate for hypothesis test with α Type I error rate is given by:

$$\beta = \Phi\left[(c - \tfrac{1}{2})\sqrt{n}\right]$$

30 Aug 2019

Computer Age Statistical Inference CHAPTER 03

→ Bayesian Inference

Bayes Rule: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

- like in Frequentist Inference, the fundamental unit of inference is a family of probability densities.

$F = \{ f_\mu(x) ; x \in X, \mu \in \Omega \}$ — x is a point in the sample space X; μ is an unobserved point in the parameter space Ω

→ we observe x from $f_\mu(x)$ & infer μ.

→ Bayesian Inference also assumes the knowledge of a prior density $g(\mu)$, $\mu \in \Omega$

e.g. if μ is the average height of a room, I would need to state a priori (before observing the sample x)

→ Bayesian works for cases where we don't have a lot of data but we have expert knowledge.

↳ in frequentist, with few data the confidence intervals would be very thick. (wide)

- Bayes Theorem (aka Bayes Rule) is a rule to combine prior knowledge in $g(\mu)$ with current evidence on x.

→ How to consistently update our belief of μ?

Let $g(\mu|x)$ denote the posterior density of μ

↳ after observing the data vector x

Bayes Rule → posterior distribution

$$g(\mu|x) = \frac{g(\mu) \cdot f_\mu(x)}{f(x)}$$ where $f(x)$ is the marginal density of x

$$f(x) = \int_\Omega f_\mu(x) \cdot g(\mu)\, d\mu$$

- $g(\mu)$ → prior distribution
- $f_\mu(x)$ → likelihood that μ is the true parameter given x
- $f(x)$ → probability of having observed the vector (data) x after averaging across all the possible values that μ can take.

Frequentists assumes that there is a true distribution where the data comes from. Also assumes that just like the sample x you got, there are an infinite number of them

In Bayesian, x is fixed (only one sample)
The only thing that changes is the belief of μ

Bayes Rule can be written as:

$$g(\mu|x) = c_x\, L_x(\mu)\, g(\mu) \quad \text{where } f_\mu(x) = L_x(\mu) = f(x|\mu)$$

- c_x → constant that only depends on x
  - aka. normalization constant of the posterior dist.
    - as any probability distribution integrates to 1

For any two μ_1, μ_2 values on Ω, the ratio of posterior densities is given by:

$$\frac{g(\mu_1|x)}{g(\mu_2|x)} = \frac{g(\mu_1)}{g(\mu_2)} \cdot \frac{f_{\mu_1}(x)}{f_{\mu_2}(x)}$$ ⇒ the posterior odds ratio is the prior odds ratio times the Likelihood ratio.

→ The relative probability (aka. the odds ratio) of μ_1 being the right value of the parameter μ, given the sample x

is equal to the relative odds of the prior belief times the odds ratio of the likelihood function from the observed data x.

e.g. An engineer knows she's having twins.
She asks what's the probability that they'll be identical.
The doctor says:
$\frac{1}{3}$ of twin births are identical } prior belief

Say X is a sonogram result (either same sex or opposite sex)
and same sex is observed.
(identical twins always are of the same sex, while fraternals have 0.5 probability of same or different sex)

- How does the sonogram modify our prior of 1/3 probability on the twins being identical?

$$\frac{g(I,S)}{g(F,S)} = \frac{g(I)}{g(F)} \cdot \frac{f_I(S)}{f_F(S)}$$

D = different twins.
I = identical twins.
S = same sex.
F = fraternal.

$$= \frac{1/3}{1-1/3} \cdot \frac{1}{1/2}$$

$\frac{1}{1/2} \rightarrow$ if they are identical, what's the prob. of being same sex
$\frac{1}{1-1/3} \rightarrow$ " fraternal, " same sex

$= 1$

↳ g(I,S) & g(F,S) are equaly likely.

» Fraternal & Identical are equaly likely

	S	D	joint probability distribution
I	1/3	0	$1/3 = g(I)$
F	1/3	1/3	$2/3 = 1 - \frac{1}{3} = g(F)$

e.g. The binomial distribution counts the number of successes or failures in n trials

If a random variable is distributed binomialy

$X \sim Bin(n, p) \Rightarrow P(X = k | p, n)$

↳ number of successes being k

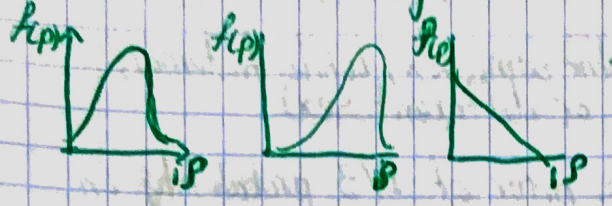
p = probability of success
n = number of observations (trials)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\cdot\Gamma(n-k+1)}$$

$P(X=k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$

- $P(X=k \mid n, p)$: Probability that the # of successes is exactly k
- p^k: observing k successes
- $(1-p)^{n-k}$: observing n-k failures

> let's assume that before observing our data, we belief p to look either of these ways:



⇒ all of these can be modeled using the Beta distributions. (family of)

So let's model our prior distribution on P as:

$$\pi(p \mid \alpha, \beta) \sim \text{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \quad ; 0 \le p \le 1$$

(π: prior)

where $B(\alpha,\beta)$ is the Beta function

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$ where Γ is the gamma function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-z}\, dx \quad ; z > 0$$

↳ is the continuous generalization of the factorial function.

$\Gamma(n+1) = n!$

$\frac{1}{B(\alpha,\beta)}$ is the normalizing constant

03 Sep 2019

[also: $B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\, dp$

> if $\pi(p \mid \alpha, \beta) \sim \text{Beta}(\alpha, \beta)$, then the mean of the Beta distribution is given by:

$$\underbrace{E(p \mid \alpha, \beta)}_{\text{The mean}} = \frac{\alpha}{\alpha+\beta}$$

For the following analysis is convenient to parameterize our Beta distribution such that the mean can be represented using only one parameter...
So, let.

$\mu = \frac{\alpha}{\alpha+\beta}$; $M = \alpha+\beta$ $\Longrightarrow$ new prior $\pi(p|\mu,M) = \text{Beta}(M\cdot\mu, M(1-\mu))$

↳ under this parametrization the expected value E is:

$E(p|\mu,M) = \mu$; $V(p|\mu,M) = \mu(1-\mu)/M+1$

↳ expected value of p ↳ variance of p

▷ Using this new prior, we can proceed to get the posterior distribution.

$P(p|k) \propto$ (aka. is proportional to) $\ell(k|p)\cdot\pi(p|\mu,M)$

↳ posterior dist. ↳ likelihood function ↳ prior

our belief about p after observing k successes

* → we are applying the Bayes rule, without the normalizing constant $1/B(\alpha,\beta)$ because it doesn't alter the distribution "shape".
- Bayes rule only needs the proportionality.

▷ Then $P(p|k) \propto (p^k(1-p)^{n-k})(p^{M\mu-1}(1-p)^{M(1-\mu)-1})$

↳ without $\binom{n}{k}$ because the same reason as *

$\propto p^{(k+M\mu)-1}(1-p)^{n-k+M(1-\mu)-1}$

$\propto \text{Beta}[k+M\mu, n-k+M(1-\mu)]$ $\Rightarrow$ probability density function

---

▷ if the prior for p is Beta & the likelihood is Binomial, then the posterior for p is also Binomial.

↳ that is a conjugate prior (aka. the posterior is of the same family as the prior)

▷ the expected value of the posterior P given we have observed k success is:

$E(P,k) = \frac{k+M\mu}{n+M}$

↑ size of the sample

$x = [\underbrace{0,0,1,0,1,\ldots,0,1}_{n}]$

$k = \sum x_i$

k = number of successes

M, μ = the two prior parameters, which is ~~(equivalent to have observed k)~~ the # of success in the past (before observing the sample)

M = is equivalent to having had a sample ~~in~~ size in the past. (aka. the ~~size~~ size of the prior sample)

μ is the expectation for p

➢ we can say our prior assumed having seen $M\mu$ successes in M trials before running the current experiment.

➢ What is our prediction of the number of successes given our posterior for p?

- Using conditional probability, we can write:

$$m(k|\mu, M, n) = \int_0^1 \ell(k|p) \cdot P(p|\mu, M, k) \cdot dp$$

↓ marginal distribution of the # of successes in n future trials, given parameters μ, M & n

$$= \frac{\Gamma(M)}{\Gamma(M\mu) \cdot \Gamma(M(1-\mu))} \binom{n}{k} \int_0^1 p^{k+M\mu-1} \cdot (1-p)^{n-k+M(1-\mu)-1} dp$$

This is the Beta function

$$= \frac{\Gamma(M)}{\Gamma(M\mu) \cdot \Gamma(M(1-\mu))} \cdot \frac{\Gamma(n+1)}{\Gamma(k+1) \cdot \Gamma(n-k+1)} \cdot \frac{\Gamma(k+M\mu) \cdot \Gamma(n-k+M(1-\mu))}{\Gamma(n+M)}$$

⇒ Probability mass function

= the probability distribution for the number of successes you are expecting to observe.

= is the posterior predicted distribution as the random variable (number of successes k) is discrete.

successes

$\alpha = k - M\mu$

$\beta = \underbrace{n-k}_{\text{failures}} + M(1-\mu)$

➢ look for "Conjugate Prior" tables for more examples.