$H_0: p \leq .5$

$H_A: p > .5$

$\sum_{i=1}^{100} X_i$ (= # of herds in 100 boxes)

$\sim Bin(100, p)$

under H_0

$\sim Bin(100, p = .5)$

Var(Bin(n,p)) $= np(1-p)$

$\sum_{i=1}^{100} X_i \sim N\left(50, \sqrt{100(.5)(.5)}\right)$

$= 5$

experiment: critical region

$\sum X_i \geq 65$

$\Rightarrow$ 3 standard deviations

level of the test $P($false positive if H_0 is true$)$

type I error

$P(\text{fix pain}) = P(Z \geq 3)$

$= 1 - \text{norm.cdf}(3)$

$\approx .0014$

let's specify

level is $\alpha = .01$

$\text{norm.ppf}(.99, \text{loc}=50, \text{scale}=5)$

≈ 61.6

$\rightarrow$ critical region is

62 heads to give up

$\therefore \alpha = .01$ test in

case

56 heads $\rightarrow$ p-value is the probability of how an at least as extreme as this under H_0

what is the smallest level α under which we would have rejected H_0

$P_{H_0}\left(\sum_{i=1}^{n} X_i \geq 56\right) = 1 - \text{norm.cdf}(56, \text{loc}=50, \text{scale}=5)$

Type II error → when you fail to reject H_0 when some H_A is actually true

power of a test (given alternative H_A that we assume is true)

↳ P(test statistic being in the critical region if H_A is true)

$p = .7$ reject if $\sum x_i \geq 62$

under this H_A

$$\sum x_i \sim N(70, \sqrt{100 \cdot .7 \cdot .3})$$

$$P_{H_A}(\sum x_i \geq 62)$$

$$= 1 - \text{norm.cdf}(62, \text{loc}=70, \text{scale}=\text{np.sqrt}(100 \cdot .7 \cdot .3))$$

2.96

$X_1, \ldots, X_{100}$

m very large #

generate m bootstrap samples

↳ take sum of each sample

if $\alpha = .01$

if the 1st percentile of

bootstrapped sum is bigger

than 50, then reject H_0

Regression (Linear Regression)

Let's suppose my data is paired

$(X_1, Y_1), \ldots, (X_n, Y_n)$

X_i's are not random - fixed

$Y_i = \alpha + \beta X_i + \varepsilon_i$ $\quad \varepsilon_i \sim N(0, \sigma)$

ε_i mutually independent

$$Y_i \sim N(\alpha + \beta X_i, \sigma)$$

we are going to estimate α, β, σ using MLE

$$L(\alpha, \beta, \sigma \mid (X_1, Y_1), \ldots, (X_n, Y_n))$$

$$= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\right)$$

$$\ell(\alpha, \beta, \sigma) = -\frac{n}{2} \log(2\pi) - n \log(\sigma)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \quad -$$

if I maximize

$$- \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

minimize

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

↳ finding the least square line

$$\frac{\partial \ell}{\partial \alpha} \quad \frac{1}{\sigma^2} \sum (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial \ell}{\partial \alpha} = \frac{1}{\sigma^2} \sum (y_i - \alpha - \beta x_i) = 0$$

$$\sum (y_i - \alpha - \beta x_i) = 0$$

$$n\hat{\alpha} = \sum y_i - \beta \sum x_i$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha) x_i = 0$$

(multiply by σ^2)

$$= \sum_{i=1}^{n} y_i x_i - \alpha \sum x_i - \beta \sum_{i=1}^{n} x_i^2$$

$$= \sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta \bar{x}) \sum x_i - \beta \sum x_i^2$$

$$= n \cdot \overline{yx} - (\bar{y} - \beta \bar{x}) n \bar{x} - \beta n \overline{x^2} = 0$$

$$\beta (\overline{x^2} - \bar{x}^2) = \overline{yx} - \bar{y}\bar{x}$$

$$\hat{\beta} = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \hat{\beta}$$

$\hat{\beta}$ is normal dist'n

$$E[\hat{\beta}] = \frac{\sum (x_i - \bar{x}) E[Y_i]}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i)}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\alpha \left(\sum x_i - \bar{x}\right) + \beta \sum x_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\sum x_i (x_i - \bar{x})$$

$$= \sum x_i^2 - n\bar{x}^2$$

$= \beta$

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Likewise, we can show that the dist'n of $\hat{\alpha}$

$$\sim N(\alpha, \ldots)$$

$$SS_x = \sum (x_i - \bar{x})^2$$

SSR = sum of squared residuals

$$= \sum_{i=1}^{n} (y_i - \underbrace{\hat{\alpha} - \hat{\beta} x_i}_{\hat{\varepsilon}_i})^2$$

$$\sigma^2 SSR \sim \chi^2_{n-2}$$

$$\hat{\beta} \pm \sqrt{\frac{SSR}{(n-2) S_{xx}}}\, t_{\alpha/2, n-2}$$

we can also bootstrap this

$(1,3), (2,5), (\ ,\ ), (4,\ )$

$(x_1, y_1), \ldots, (x_n, y_n)$

resample the pairs $\rightarrow$ generate n bootstrap samples

calculate regression coefficients for each sample

find appropriate sample

percentiles

Prediction Intervals ...

$X_1, \ldots, X_n \sim N(\mu, \sigma)$

X_{n+1} comes in the future

let's say I know μ, σ

can show that for any predictor p

$$E[(X_{n+1} - p)^2] \geq E[(X_{n+1} - \mu)^2]$$

μ is the optimal predictor

$\Rightarrow$ prediction with $\bar{X}$ $\quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$Var(X_{n+1} - \bar{X}) = E[(X_{n+1} - \bar{X}_n)^2]$$

$X_{n+1} \sim N(\mu, \sigma)$

$\bar{X}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

$Var(aX \pm bY)$

$= a^2 Var(X)$

$+ b^2 Var(Y)$

~~$+ 2ab\,Cov(X,Y)$~~

$Var(X_{n+1} - \bar{X}_n)$

$$= V_{\sigma}(X_{n+1}) + V_{\sigma}(\bar{X}_n)$$

$$= \sigma^2 + \frac{\sigma^2}{n}$$

$$= \sigma\left(1 + \frac{1}{n}\right)$$

⟹ prediction error: part of error
coming from estimating $E[X_{n+1}]$
rest coming from $V_{\sigma}(X_{n+1})$

⟹ prediction errors in regression models

(X_{n+1}, Y_{n+1})

Y_{n+1} I don't

choose X_{n+1}

$$Y_{n+1} \sim N(\alpha + \beta x_i, \sigma)$$

⟹ calculate a prediction interval for Y_{n+1}?

use the bootstrap to generate prediction intervals