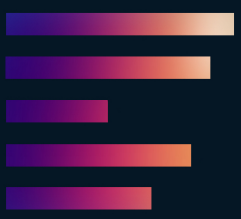
SPRING 2026

MATH 450: Mathematical Statistics



Monday, January 12

- question:
  - Suppose we have a sample (n iid random variables $X_1, \dots, X_n$) from a known distribution with unknown parameter θ.
  - How can we approximate θ?

- example: 50 values from $\mathcal{U}(0, \theta)$ (θ unknown)
  - A few possibilities:
    1. $\max\{X_i\}$
    2. $2\bar{X}$
    3. $n \cdot \min\{X_i\}$

    (all sample dependent)

- definition: statistic
  - A statistic is a function of a sample $X_1, \dots, X_n$.

- example: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- definition: estimator
  - An estimator is a statistic used to estimate a parameter.

- example: $\bar{X}$ estimates μ.

- topic: desirable properties of estimators
  1. unbiased: $E(\hat{\theta}) = \theta$ ($\hat{\theta}$: estimator (statistic))
     - example: $\bar{X}$ is an unbiased estimator of μ.

       proof
       - $E(\bar{X}) = E(\frac{1}{n}\sum X_i)$
         $= \frac{1}{n} E(\sum X_i)$
         $= \frac{1}{n} \sum E(X_i)$
         $= \frac{1}{n} \cdot n\mu$
         $= \mu$
     - definition: bias
       - the bias of a statistic is $B = E(\hat{\theta}) - \theta$
  2. low variance: $\mathrm{Var}(\hat{\theta})$ ("efficiency")
  3. use all possible information: "sufficiency"
     - example: $\bar{X}$ and X_1 both estimate μ
  4. limit behavior: as $n \to \infty$, $\hat{\theta} \to \theta$ [$\mathrm{Var}(\hat{\theta}) \to 0$] ("consistency")

Monday, January 12

- example: show $S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

  $\sigma^2 = E(X^2) - E(X)^2$, $\mathrm{Var}(X)$

  - $E(S^2) = \frac{1}{n-1} E(\sum (X_i - \bar{X})^2)$
    $= \frac{1}{n-1} E[\sum X_i^2 - 2\bar{X}\underbrace{\sum X_i}_{n\cdot\bar{X}} + \underbrace{\sum (\bar{X})^2}_{n\cdot(\bar{X})^2}]$
    $= \frac{1}{n-1} E[\sum X_i^2 - n(\bar{X})^2 + n(\bar{X})^2]$
    $= \frac{1}{n-1} E(\sum X_i^2 - n(\bar{X})^2)$
    $= \frac{1}{n-1} [\sum E(X_i^2) - n \cdot E(\bar{X}^2)]$
    $= \frac{1}{n-1} [\sum (\sigma^2 + \mu^2) - n(\mathrm{Var}(\bar{X}) - E(\bar{X})^2)]$
    $= \frac{1}{n-1} [n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n} + \mu^2)]$
    $= \frac{1}{n-1} [n\sigma^2 - \sigma + n\mu - n\mu^2)$
    $= \frac{1}{n-1} [(n-1)\sigma^2]$
    $= \sigma^2$

Wednesday, January 14

- topic: the method of moments (MoM)
  - idea:
    1. find theoretical mean.
    2. set it equal to observed mean
  - example: Suppose $X_1, \dots, X_{50} \sim \mathcal{U}(0, \theta)$
    - $\mu = E(X) = \frac{\theta}{2}$
    - $\mu = \frac{\theta}{2} = \bar{X}$
      $\Rightarrow \hat{\theta} = 2\bar{X}$
  - example: Suppose $X_1, \dots, X_{50} \sim \Gamma(\alpha, \theta)$.
    1. compute 1st two theoretical moments (2 unknowns $\Rightarrow$ 2 moments).
    2. set them equal to 1st two observed moments

    1. $E(X) = \alpha\theta$
       $\mathrm{Var}(X) = E(X^2) - E(X)^2$
       $\Rightarrow E(X^2) = \mathrm{Var}(X) + E(X)^2$
       $= \alpha\theta^2 + \alpha^2\theta^2$
       $= \alpha\theta^2(\alpha + 1)$

    2. $E(X) = \alpha\theta = \frac{1}{n}\sum X_i$
       $E(X^2) = \alpha\theta^2(\alpha+1) = \frac{1}{n}\sum X_i^2$
       Let $\frac{1}{n}\sum X_i = m_1$, Let $\frac{1}{n}\sum X_i^2 = m_2$
       $\alpha\theta = m_1 \Rightarrow \alpha = \frac{m_1}{\theta}$
       $\alpha\theta^2(\alpha+1) = m_2$
       $\frac{m_1}{\theta} \cdot \theta^2(\frac{m_1}{\theta} + 1) = m_2$
       $m_1^2 + m_1\theta = m_2$
       $\Rightarrow \hat{\theta} = \frac{m_2 - m_1^2}{m_1}$ & $\hat{\alpha} = \frac{m_1}{\theta} = \frac{m_1^2}{m_2 - m_1^2}$

topic: maximum likelihood estimation (MLE)

example: Let $X_1, \dots, X_n$ be Bernoulli RVs. $X_i \sim Bin(1, p)$.
Estimate p (ie., find $\hat{p}$).

idea:
- For a sample $x_1, \dots, x_n$, pick $\hat{p}$ that gives the greatest likelihood of the observed sample.
- We need the joint pmf of $X_1, \dots, X_n$.
- Each individual pmf is

$$f(x) = \begin{cases} p & , x = 1 \\ 1-p & , x = 0 \end{cases} = p^x (1-p)^{1-x} \quad , x = 0, 1$$

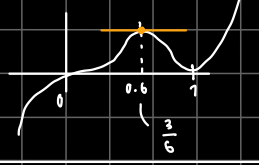
- $f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$
  $= P(X_1 = x_1) P(X_2 = x_2) \dots P(X_n = x_n)$
  $= f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n)$
  $= \prod f(x_i)$
  $= \prod p^{x_i} (1-p)^{1-x_i}$
  $= p^{\sum x_i} (1-p)^{n - \sum x_i}$

shift in perspective from x to p

"likelihood function" $L(p)$

not a pmf of p

example: Let $n = 6$, $x_i = 1, 0, 0, 1, 1$
- $L(p) = p^3 (1-p)^2$



Friday, January 16

recall:
- $X_i \sim B(1, p)$
- Want to estimate p using a sample size of n
- $L(p) = p^{\sum x_i} (1-p)^{n - \sum x_i}$

continuing:
- Goal: Maximize $L(p)$ on $[0, 1]$
- $L'(p) = \sum x_i \, p^{\sum x_i - 1} (1-p)^{n - \sum x_i} + p^{\sum x_i} (n - \sum x_i) (1-p)^{(n - \sum x_i) - 1} (-1)$
  $0 = p^{\sum x_i - 1} (1-p)^{n - \sum x_i - 1} \left[\sum x_i (1-p) - p(n - \sum x_i) \right]$
- Assume $p \neq 0$ and $p \neq 1$
  $\Rightarrow 0 = \sum x_i (1-p) - p(n - \sum x_i)$
  $= \sum x_i - p \sum x_i - np + p \sum x_i$
  $\Rightarrow np = \sum x_i \Rightarrow \hat{p} = \frac{1}{n} \sum x_i$ (success = proportion of successes)
  ↳ maximum likelihood estimator (MLE)
- If $p = 0$ or $p = 1$, then
  $L(p) = 0$

---

fact: If $L(\theta) \neq 0$, then TFAE:
- $L'(\theta) = 0$
- $\frac{d}{d\theta} \log L(\theta) = 0$

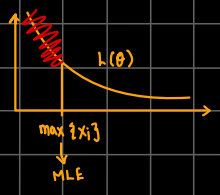
proof:
- $\frac{d}{d\theta} \log L(\theta) = \frac{L'(\theta)}{L(\theta)}$ ⊠

$L(\theta) = \prod_{i=1}^{n} f(x_i)$ ← "log likelihood"
$\Rightarrow \log L(\theta) = \sum \log f(x_i)$

example: 6.4-5 (a)
- $f(x; \theta) = \frac{1}{\theta^2} x e^{-\frac{x}{\theta}}$, $x > 0$, $\theta > 0$
- $L(\theta) = \prod \frac{1}{\theta^2} x_i e^{\frac{-x_i}{\theta}}$
- $\log L(\theta) = \sum \log \left(\frac{1}{\theta^2} x_i e^{\frac{-x_i}{\theta}} \right)$
  $= \sum (\log 1 - \log \theta^2) + \log x_i + \frac{x_i}{\theta}$
  $= \sum [\log x_i - 2 \log \theta + x_i/\theta]$
- $\frac{d}{d\theta} \log L(\theta) = \sum [\theta^{-1} \cdot \frac{2}{\theta} + x_i \theta^{-2}]$
- $0 = \sum (-2/\theta + x_i \theta^{-2})$
  $= -2n/\theta + 1/\theta^2 \sum x_i$
  $2n = \frac{1}{\theta} \sum x_i$
  $\Rightarrow \theta = \frac{1}{2n} \sum x_i$
  $= \frac{1}{2} \bar{x}$ ⊠

---

example: $x_i \sim \mathcal{U}(0, \theta)$
- $f(x; \theta) = 1/\theta$
- $f(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta}$
  $L(\theta) = \theta^{-n}$
- maximize on $[\max(x_i), \infty]$



Wednesday, January 21

- example:
  - Suppose $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$
  - Find MLE for μ & σ^2.
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
  - $L(\mu,\sigma^2) = \prod f(x_i)$
    $= \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$
    $= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{\sum -(x_i-\mu)^2/2\sigma^2}$
  - $\log L(\mu,\sigma^2) = \log\left(2\pi\sigma^2\right)^{-n/2} + \sum -(x_i-\mu)^2/2\sigma^2$
    $= -n/2 \log 2\pi\sigma^2 - 1/2\sigma^2 \sum (x_i-\mu)^2$
  - $\frac{\partial}{\partial\mu} \log L(\mu,\sigma^2) = 1/\sigma^2 \sum (x_i-\mu) = 0$
    $\Rightarrow \sum x_i = \sum \mu = n\mu.$ ⊠
  - $\frac{\partial}{\partial\sigma^2} \log L(\mu,\sigma^2) = -n/2\sigma^2 - 1/(2\sigma^2)^2 \sum (x_i-\mu)^2 = 0$
    $\Rightarrow n/2\sigma^2 = 1/(2\sigma^2)^2 \sum (x_i-\mu)^2$
    $\Rightarrow \sigma^2 = 1/n \sum (x_i-\bar{x})^2$ ⊠
    ↑ biased estimator (underestimates)

$\frac{d}{d\sigma^2}\left[\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right] = (-2\sigma^2)^{-2}\sum(x_i-\mu^2) = -\frac{1}{(2\sigma^2)^2}\sum(x_i-\mu^2)$

$\frac{d}{d\sigma^2}\left[\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right]$

$\frac{d}{d\sigma^2}\left[-\frac{n}{2}\log 2\pi\sigma^2\right]$

$\frac{d}{d\sigma^2}\left[-\frac{n}{2}(\log 2\pi + \log\sigma^2)\right]$

$\frac{d}{d\sigma^2}\left[-\frac{n}{2}\log\sigma^2\right] = -\frac{n}{2\sigma^2}$

- definition: sufficiency (computational)
  - Let $Y = u(X_1, \dots, X_n)$ be a statistic for a distribution with parameter θ.
  - Then, Y is sufficient for θ if the joint pdf can be written in a factored form

$$f(x_1, \dots, x_n; \theta) = \underbrace{\phi(y;\theta)}_{\text{no dependence on the sample except } y} \cdot \overbrace{h(x_1, \dots, x_n)}^{\text{no dependence on } \theta,\ f(x_i)}$$

- example:
  - Let $X_1, X_2, \dots, X_n \sim \text{Pois}(\lambda)$
  - $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
  - $f(x_1, \dots, x_n) = \prod \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-n\lambda} \prod \frac{\lambda^{x_i}}{x_i!}$
    $= \underbrace{e^{-n\lambda}\lambda^{\sum x_i}}_{f(x_i,\lambda)} \cdot \underbrace{1/\prod x_i!}_{f(x_i)}$
  - ⇒ as a result, $\sum X_i$ is sufficient for λ.

Monday, January 26

- fact: sufficiency (intuitional)
  - If $Y = u(X_1, \dots, X_n)$ is sufficient for θ, then
    $P(X_1 = x_1, \dots, X_n = x_n \mid Y = y)$
    does not depend on θ.

- definition: the exponential family of distributions
  - these are the distributions whose pmfs or pdfs can be written in this form:

$$f(x;\theta) = \exp\left[K(x)\,p(\theta) + S(x) + q(\theta)\right]$$

  - pretty much all of our favorite distributions are in this family.

- example: poisson: $\text{pois}(\lambda)$
  - $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{e^{\ln(\lambda^x e^{-\lambda})}}{e^{\ln(x!)}}$
    $= \frac{e^{x\ln\lambda - \lambda}}{e^{\ln(x!)}}$
    $= e^{x\ln\lambda - \lambda - \ln(x!)}$
    $\exp(x\ln(\lambda) - \lambda - \ln(x!))$
    $K(x)$ $p(\lambda)$ $q(\lambda)$ $S(x)$

- example: chi-squared: $\chi^2(r)$
  - $f(x;r) = \frac{x^{r/2-1} e^{-x/2}}{\Gamma(r/2)\,2^{r/2}} = \frac{e^{\ln(x^{r/2-1}\cdot e^{-x/2})}}{e^{\ln(\Gamma(r/2)\cdot 2^{r/2})}}$
    $= \frac{e^{(r/2-1)\ln(x)} + e^{\ln(e^{-x/2})}}{e^{\ln(\Gamma(r/2)\cdot 2^{r/2})}}$
    $= e^{(r/2-1)\ln(x)} + e^{-x/2} - e^{\ln(\Gamma(r/2)\cdot 2^{r/2})}$
    $= \exp\left[(r/2-1)\ln x - x/2 - \ln\left[\Gamma(r/2)\cdot 2^{r/2}\right]\right]$
    $p(r)$ $K(x)$ $S(x)$ $q(r)$

---

fact: exponential family sufficiency

- for such a distribution, the statistic
  $Y = \sum k(x_i)$
  is sufficient for θ.

proof:

- if $f(x) = \exp[k(x)p(\theta) + S(x) + q_1(\theta)]$
  and $x_1, \dots, x_n$ are iid from this distribution.

- $f(x_1, \dots, x_n; \theta) = \prod f(x_i)$
  $= \prod \exp[k(x_i)p(\theta) + S(x_i) + q_1(\theta)]$
  $= \exp[np(\theta)\sum k(x_i) + \sum S(x_i) + nq_1(\theta)]$
  $= \exp[np(\theta)\sum k(x_i) + nq_1(\theta) + \sum S(x_i)]$
  $= \exp[np(\theta)\sum k(x_i) + nq_1(\theta)] \cdot \exp(\sum S(x_i))$
  $f(x_i, \theta)$ ↘ $h(x_i)$

---

fact:

- if $Y = u(X)$ is sufficient for θ and $W = g(Y)$ is another statistic where g is invertible, then W is also sufficient.

$u: X \to Y$, $g: Y \to W$ } $g \circ u : X \xrightarrow{u} Y \xrightarrow{g} W$

Wednesday, January 28

thm: Rao-Blackwell

- suppose $Y_1 = u_1(X_1, \dots, X_n)$ is sufficient for some parameter θ and $Y_2 = u_2(X_1, \dots, X_n)$ is an unbiased estimator of θ.
- then $E(Y_2 \mid Y_1 = y_1)$ describes an unbiased sufficient estimator of θ.

---

topic: confidence intervals

- point estimators are <u>always</u> wrong.
- all we can hope for in the real world is that they're <u>close</u>.
- a practical approach is to give an <u>interval estimate</u>.
- a <u>level C confidence interval</u> for a parameter θ is an interval (a, b) computed from a sample such that:

$$P(a(x_i) < \theta < b(x_i)) = C = 1 - \alpha$$

(statistics; how often the parameter is in the interval; how often it isn't in the interval)

example:

- suppose $X_i \sim N(\mu, \sigma^2)$ (unknown)
- create a level C confidence interval for μ.
- <u>recall:</u> $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$
  (stand-in for standard normal when standard distribution is unknown)

1. find $\bar{x}$
2. find s
3. find t_*
4. find E
5. lower = $\bar{x} - E$, upper = $\bar{x} + E$

- $-t_* = qt(.25, n-1)$
  $t_* = qt(.975, n-1)$

- let's choose to get an interval symmetric about $\bar{X}$, one of the form:

$$\mu = \bar{x} \pm E$$

$$P(|\bar{x} - \mu| < E) = C$$

$$P\left(\left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| < \frac{E}{S/\sqrt{n}}\right) = C$$

($\frac{\bar{x} - \mu}{S/\sqrt{n}}$ ↙ T)

$$P(-t_* < T < t_*) = C$$

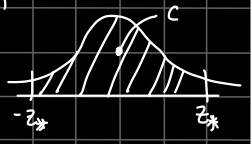
$$t_* = \frac{E}{S/\sqrt{n}}$$

$$\Rightarrow E = t_* \cdot S/\sqrt{n}$$

Friday, January 30

topic: confidence intervals for proportions

- suppose $X_1, X_2, \dots, X_n$ are iid from $\text{Bin}(1,p)$.
- we want to estimate p.
- $\hat{p} = \frac{1}{n}\sum X_i$ (MLE) is unbiased and sufficient for p.
- we know that $\mu_{X_i} = p$, $\sigma^2_{X_i} = p(1-p)$
- By the CLT, $\hat{p} \sim N(\mu = p, \sigma^2 = 1/n \cdot p(1-p))$
- $\frac{\bar{x} - \mu}{\sigma} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
- $P\left(\left|\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| < z_*\right) = C$



- $P\left(|\hat{p} - p| < z_*\sqrt{p(1-p)/n}\right)$
- $-z_*\sqrt{p(1-p)/n} < \hat{p} - p < z_*\sqrt{p(1-p)/n}$
- $\hat{p} - z_*\sqrt{p(1-p)/n} < p < \hat{p} + z_*\sqrt{p(1-p)/n}$

(p unknown; p unknown)

idea 1: in the formula for variance only, approximate p with $\hat{p}$. (Wald interval). (tends to under-cover)

- margin of error $= z_*\sqrt{\hat{p}(1-\hat{p})/n}$ (z_*: qnorm)

idea 2: solve for p directly. (Wilson interval)

- find endpoints by changing "<" to "="

$|\hat{p} - p| = z_x\sqrt{p(1-p)/n}$

$(\hat{p} - p)^2 = z_x^2 \cdot 1/n \cdot p(1-p)$

$\hat{p}^2 - 2p\hat{p} + p^2 = z_x^2 \cdot 1/n \cdot p - p^2$

$0 = 1/n \cdot z_*^2(p - p^2) - \hat{p}^2 - 2p\hat{p} + p^2$

$= p^2(1 + z_*^2/n) - p(2\hat{p} + z_*^2/n) + \hat{p}^2$

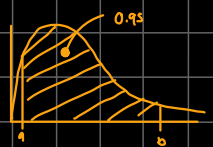
$p = \frac{(2\hat{p} + z_*^2/n) \pm \sqrt{(2\hat{p} + z_*^2/n) - 4(1 + z_*^2/n)\hat{p}}}{2(1 + z_*^2/n)}$

prop.test(x, n, conf.level, correct = FALSE) (successes, trials; Yates' correction)

Monday, February 2

example:

- Find c & d such that $P(c < \sigma^2 < d) = 0.95$ when $X_i \sim N(\mu, \sigma^2)$, $i = 1,2,\dots,40$



a = qchisq(0.025, 39)
b = qchisq(0.975, 39)

- We know that $39S^2/\sigma^2 \sim \chi^2(39)$

(not always true)

generally $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$

- $P(a < 39S^2/\sigma^2 < b) = 0.95$

$P(1/a > \sigma^2/39S^2 > 1/b) = 0.95$

$P(39S^2/a > \sigma^2 > 39S^2/b) = 0.95$

$\Rightarrow P(39S^2/b < \sigma^2 < 39S^2/a) = 0.95$

$P(0.6715S^2 < \sigma^2 < 1.6495S^2) = 0.95$

---

topic: hypothesis testing

- so far:
  - point estimates $\hat{\theta}$ for parameter θ
    - MLE and MOM estimators
    - desirable properties (unbiased & sufficient)
  - interval estimates.
- next up:
  - hypothesis testing
    - we don't know θ but we do have a value θ_0 of interest.
- idea: (H_0: null hypothesis → $H_0: \theta = \theta_0$)
  - assume H_0 is true then compute the probability of a sample "like the one you got".
  - if this probability (p-value) is low, H_0 is not plausible.
  - a sample $X_1, X_2, \dots, X_n$ defines a point in $\mathbb{R}^n$.
  - one way of defining a test is by choosing which samples will cause us to reject H_0.

---

definition: critical region (given we accept the null, what are the chances we still land in the rejection area? we want that prob $< \alpha$.)

- A critical region C (or rejection region) of size α is a subset of $\mathbb{R}^n$ such that

$$P((X_1, \dots, X_n) \in C \mid H_0) = \alpha$$ ($\theta = \theta_0$; false positive rate)

---

example: ($H_A: \mu \neq 3.6$)

- $H_0: \mu = 3.6$ where $X_i \sim N(\mu, 1.2^2)$, $n = 272$
- $Z = \frac{\bar{x} - \mu}{\sigma^2}$

$= \frac{\bar{x} - 3.6}{1.2/\sqrt{272}} = \frac{3.488 - 3.6}{1.2/\sqrt{272}}$

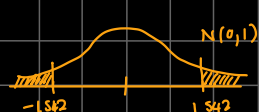
$= -1.542$

(two-tailed test)

- $p = 2 \times \text{pnorm}(z)$

$= 0.123$ is large so we don't reject.

ex: $\alpha = 0.05$



- alternatively, we can plan on rejecting H_0 if

$$\left| \frac{\bar{X} - 3.6}{1.2/\sqrt{272}} \right| > 1.96$$

- since $P(|Z| > 1.96) = 0.05$, this defines a critical region of size $\alpha = 0.05$.

Wednesday, February 4

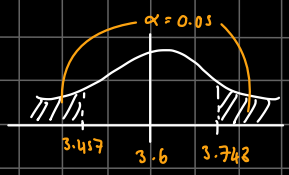
recall:

- $P[(X_1, \dots, X_n) \in C \mid \theta = \theta_0] = \alpha$
- our perspective: choosing C defines the test.

---

example:

- $X_1, \dots, X_{272} \sim N(\mu, 1.2^2)$
- $H_0: \mu = 3.6$

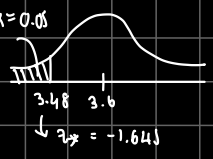


- Let $C_1 = \{\bar{X} < 3.457 \text{ or } \bar{X} > 3.743\}$
- this has size $\alpha = 0.05$

$$P(\bar{X} \in C_1 \mid \mu = 3.6) = 0.05$$

- $\text{pnorm}(3.457, 3.6, 1.2/\sqrt{272}) \approx 0.025$
- $1 - \text{pnorm}(3.743, 3.6, 1.2/\sqrt{272}) \approx 0.025$

0.05

- 

$z_* = -1.645$

in $N(3.6, 1.2^2)$, cutoff $= 3.6 - 1.645\,(1.2/\sqrt{272}) \approx 3.48$

- Let $C_2 = \{\bar{X} < 3.48\}$

- Let $C_3 = \{X_1 < \text{qnorm}(0.05, 3.6, 1.2)\}$

want $P(X_1 < z \mid \mu = 3.6) = 0.05$

$C_3 = \{X_1 < 1.63\}$

- Let $C_4 = \{\max\{X_i\} < 6.36\}$

- So far we've only considered type I errors (false negative) when describing critical regions. In the above example, $\alpha = 0.05$ for each region.
- we'd also like to limit β, the probability of a type II error:

$$\beta = P(\text{don't reject} \mid H_0 \text{ is false})$$

(false positive)

$$\beta(\theta) = P(\bar{X} \notin C \mid \theta \neq \theta_0)$$

---

definition: power function of a test, $K(\theta)$

- the power function of a test

$H_0: \theta = \theta_0$

with rejection region C, is

$K(\theta) = 1 - \beta(\theta)$

power = $P(\text{reject } H_0 \mid H_0 \text{ is false})$ (true positive)

$K(\theta) = \mathbb{P}(\bar{X})$

= probability of detecting a true positive when the actual parameter value is θ

---

example:

- $X_1, \dots, X_{34} \sim \text{bin}(1, p)$.
- we want to test

$H_0: p = 0.25$

- Let $C = \{\sum X_i \leq 4\}$

$\alpha = P(\text{reject} \mid H_0 \text{ is true})$

$= P(\sum X_i \leq 4 \mid p = 0.25)$

$= P(Y \leq 4)$ in $\text{bin}(34, 0.25)$

- $K(\theta) = P(\text{reject} \mid p)$

$= P(Y \leq 4)$ in $\text{bin}(34, p)$

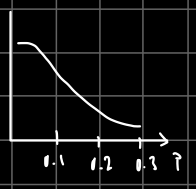
- $K(0.1) = \text{pbinom}(4, 34, 0.1)$

$= 0.75$

- $K(0.2) = \text{pbinom}(4, 34, 0.2)$

$= 0.16$

- $K(0.25) = 0.05$



Friday, February 6

example:

- $H_0: \mu = 100$
- $X_i \sim N(\mu, 10^2)$, $n = 25$
- $\alpha = \{(x_1, x_2, \dots, x_n) : \bar{X} < c\}$

- Find c such that $\alpha = 0.05$ (false positive rate)
  - $\alpha = P(\text{reject} \mid H_0 \text{ is true})$

$0.05 = P(\bar{X} < c)$ in $N(100, 10^2/25)$ (4)

$\Rightarrow c = \text{qnorm}(0.05, 100, 2)$

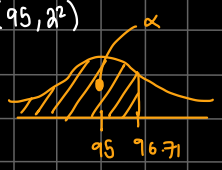
≈ 96.71

- Now, imagine this is a study on LDL cholesterol and a medication. Without the med, $\mu = 100$
- Does the new med reduce this?
- Researchers will only consider the difference important if the means is at least 5 points lower (effect size of 5).

- What is the probability of detecting a true positive?

$$K(95) = P(\text{reject } H_0 \mid \mu = 95)$$
$$= P(\bar{X} < 96.71) \text{ in } N(95, 8^2)$$
$$= \text{pnorm}(96.71, 95, 8)$$
$$= 0.804$$

↳ true positive rate



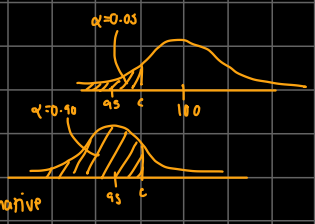
- Similarly

$$\beta(95) = 1 - K(95)$$
$$= 0.196$$

- Often, we write

$H_0 : \mu = 100$

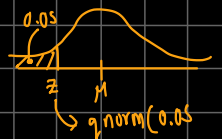
$H_1 : \mu = 95$ ↳ simple alternative



- Now, suppose we need a study with $\alpha = 0.05$ and $\beta(95) = 0.10 \Rightarrow K(95) = 0.90$.

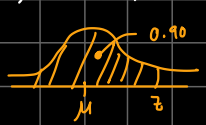
- Find c & n.
  - Using α,
    - $\alpha = P(\bar{X} < c \mid \mu = 100)$
    - $0.05 = P(\bar{X} < c)$ in $N(100, 10^2/n)$
    - $= P(Z < (c-100)/(10/\sqrt{n}))$ in $N(0,1)$

$$\Rightarrow (c-100)/(10/\sqrt{n}) = \text{qnorm}(0.05)$$
$$= -1.645 \quad (1)$$



  - Using $\beta(95)$
    - $\beta(95) = P(\bar{X} < c \mid \mu = 95)$
    - $0.90 = P(\bar{X} < c)$ in $N(95, 10^2/n)$
    - $= P(Z < (c-95)/(10/\sqrt{n}))$ in $N(0,1)$

$$\Rightarrow (c-95)/(10/\sqrt{n}) = \text{qnorm}(0.90)$$
$$= 1.282 \quad (2)$$



  - Equate and solve:

(1) $c = (-1.645) \cdot 10/\sqrt{n} + 100$

(2) $c = (1.282) \cdot 10/\sqrt{n} + 95$

$\Rightarrow n = 34.3 \approx 35$

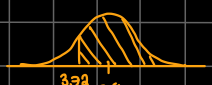
$\Rightarrow c = 97.19$ ☒

Monday, February 9

- example:
  - Let $X_1, \ldots, X_{20}$ be a sample from $N(\mu, 1.8^2)$.
  - $H_0 : \mu = 3.6$, $\alpha = 0.05$.
  - $C_1 = \{\bar{X} \geq 3.720\}$

    $C_2 = \{X_{17} > 5.574\}$
  - note:
    - under H_0, $P(\vec{X} \in C_1) = P(\vec{X} \in C_2) = \alpha = 0.05$
  - we can see that C_1 is preferable using power.
  - Consider

    $H_1 : \mu = 3.8$



  - For C_1, $K(3.8) = P(\bar{X} > 3.720 \mid \mu = 3.8)$

$$= 1 - P(X < 3.720) \text{ in } N(3.8, 1.8^2/20)$$
$$= 1 - \text{pnorm}(3.720, 3.8, 1.8/\sqrt{20})$$
$$\approx 0.864$$

  - For C_2, $K(3.8) = P(X_{17} > 5.574 \mid \mu = 3.8)$

$$= 1 - P(X < 5.574) \text{ in } N(3.8, 1.2)$$
$$= 1 - \text{pnorm}(5.574, 3.8, 1.2)$$
$$\approx 0.0696$$

- definition: best critical region
  - If C_1 and C_2 are critical regions of size α against $H_0 : \theta = \theta_0$.
  - Then C_1 is more powerful against a simple alternative $H_1 : \theta = \theta_1$ if $P(\vec{X} \in C_1 \mid H_1) > P(\vec{X} \in C_2 \mid H_1)$.
  - We say C_1 is best (or more powerful) if

$$P(\vec{X} \in C_1 \mid H_1) \geq P(\vec{X} \in D \mid H_1)$$

    for all critical regions D of size α.
  - As a rule, points in C_1 would:
    1. when H_0 is true, $L(\theta_0)$ is small in C.
    2. when H_1 is true, $L(\theta_1)$ is larger in C. (H_0 is false)
  - One way of capturing these ideas is with $L(\theta_0)/L(\theta_1)$ and should be small for all points in a best critical region.

- intuition:
  - build a best critical region from scratch starting with points where $L(\theta_0)/L(\theta_1)$ is smallest.

- thm: Neyman-Pearson Lemma
  - Let C be a critical region of size α s.t.
    1. $L(\theta_0)/L(\theta_1) \leq k$ for $(x_1, \ldots, x_n) \in C$
    2. $L(\theta_0)/L(\theta_1) \geq k$ for $(x_1, \ldots, x_n) \in C'$
  - If such a k exists, then C is a best critical region of size α.

Monday, February 16

thm: Neyman-Pearson Lemma

- Let $X_1, \dots, X_n$ be a sample to test with
  $H_0: \theta = \theta_0$
  $H_1: \theta = \theta_1$.
- Let C be a size α critical region.
- Suppose $\exists k \in \mathbb{R}$ such that

$$L(\theta_0)/L(\theta_1) < k \text{ for } (x_1, \dots, x_n) \in C$$
$$L(\theta_0)/L(\theta_1) \geq k \text{ for } (x_1, \dots, x_n) \in C'$$

- Then C is a best critical region.

proof:

- Let C, D be critical regions of size α.
- WWTS: $P(\vec{X} \in C \mid H_1) \geq P(\vec{X} \in D \mid H_1)$ (1)
  (reject, reject)
- If C & D are size α:

$$\Rightarrow \alpha = \int_C f(x_1, \dots, x_n; \theta_0) \overset{\text{notation}}{=} \int_C L(\theta_0) \quad (2)$$

$$\Rightarrow \alpha = \int_D L(\theta_0) \quad (3)$$

- WWTS:

$$\int_C L(\theta_1) \geq \int_D L(\theta_1)$$

- Using (2) and (3):

$$\int_C L(\theta_0) - \int_D L(\theta_0) = 0$$

$$= \left[\int_{C \cap D} L(\theta_0) + \int_{C \cap D'} L(\theta_0)\right] - \left[\int_{D \cap C} L(\theta_0) + \int_{D \cap C'} L(\theta_0)\right]$$

($C \cap D'$: in C; $D \cap C'$: in C')

- directly from NP condition

$$0 \leq k \int_{C \cap D} L(\theta_0) - k \int_{C' \cap D} L(\theta_1)$$

$$0 \leq k \left[\int_{C \cap D'} L(\theta_1) + \int_{C \cap D} L(\theta_1) - \int_{C \cap D} L(\theta_1) - \int_{C' \cap D} L(\theta_1)\right]$$

$$0 \leq \int_C L(\theta_1) - \int_D L(\theta_1)$$

Wednesday, February 18

example: Let $X_1, \dots, X_n$ be a sample from $N(\mu, \sigma^2)$ to test $H_0: \mu = 20$, $H_1: \mu = 22$. Show $C = \{\bar{x} > c\}$ is a best critical region for a appropriate choice of c.

- We'll compute $L(H_0)/L(H_1) = L(20)/L(22)$, set it $< k$ and solve for $\bar{x}$.
- Recall,

$$f(x) = 1/\sqrt{2\pi\sigma^2} \exp\left[-(x-\mu)^2/2\sigma^2\right]$$

- $L(\mu, \vec{x}) = \prod f(x_i)$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-1/2\sigma^2 \sum (x_i - \mu)^2\right]$$

$$\Rightarrow \frac{L(20)}{L(22)} = \frac{(2\pi\sigma^2)^{-n/2}}{(2\pi\sigma^2)^{-n/2}} \cdot \exp\left[-1/2\sigma^2 \sum (x_i-20)^2 - (x_i-22)^2\right] < k$$

$$= \exp\left[-1/2\sigma^2 \sum x_i^2 - 40x_i + 400 - x_i^2 + 44x_i - 484\right] < k$$

$$= \exp\left[-1/2\sigma^2 \sum 4x_i - 84\right] < k$$

$$= 1/2\sigma^2 \sum 84 - 4x_i < \ln k$$

$$= 1/2\sigma^2 \left(84n - 4\sum x_i\right) < \ln k$$

$$= \sum x_i > \left[84n - 2\sigma^2 \ln(k)\right]/4$$

$$\Rightarrow \bar{x} > 84n/4n - 2\sigma^2 \ln(k)/4n$$

$$\Rightarrow \bar{x} > \underbrace{21 - \frac{\sigma^2 \ln(k)}{2n}}_{c} \quad \text{(} k \text{ depends on } \alpha\text{)} \quad \blacksquare$$

$$\Rightarrow \bar{x} > c$$

- By the Neyman-Pearson Lemma, $C = \{\bar{x} > c\}$ is a best critical region for the test.

---

definition: uniformly best critical region for a composite alternative

- we say a region C in a test for a composite alternative is uniformly best if it's best against every simple alternative within the composite.

---

example: Find a uniformly best critical region for $H_0: p = p_0$, $H_a: p > p_0$ for a sample of size n, $x_i \sim \text{bin}(1, p)$.

- Let $p_1 > p_0$ be any simple alternative $H_1: p = p_1$.
- Then,

$$f(x) = p^x (1-p)^{1-x}$$
$$\Rightarrow L(x, p) = p^{\sum x_i} (1-p)^{n - \sum x_i}$$

- Now,

$$k \geq \frac{L(p_0)}{L(p_1)} = \frac{p_0^{\sum x_i} (1-p_0)^{n-\sum x_i}}{p_1^{\sum x_i} (1-p_1)^{n-\sum x_i}}$$

- Let $y = \sum x_i$

$$\Rightarrow k \geq \frac{L(p_0)}{L(p_1)} = \frac{p_0^y (1-p_0)^{n-y}}{p_1^y (1-p_1)^{n-y}}$$

Friday, February 20

topic: linear regression (fixed x_i, random Y_i)

- Suppose we have observations (x_i, Y_i) where we view the x_i's as fixed and the Y_i's as random.
- In a linear model

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$$

- Equivalently (systematic component):

$$M(x) = \beta_0 + \beta_1 x$$

(random component):

$$Y_i = \beta + \beta_1 x_i + \varepsilon_i$$

- We additionally assume

$$\varepsilon_i \sim N(0, \sigma^2)$$

(σ^2: fixed, typically unknown and same $\forall x$)

- assumptions: L I N E
  - L: linear
  - I: iid
  - N: $\varepsilon \sim N$
  - E: σ^2 is fixed $\forall x$ & equal
- Now,
  - $L(\beta_0, \beta_1, \sigma^2) = \prod f(y_i; \beta_0, \beta_1, \sigma^2)$
  - $\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$
  - $E[Y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1$

$$\Rightarrow Y_i \sim N(\beta_0 + \beta_1, \sigma^2)$$

$$\Rightarrow L(\beta_0, \beta_1, \sigma^2) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[\frac{-\sum[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right]$$

- After solving,
  - $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - $\hat{\sigma}^2 = \frac{1}{n}\sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$ (residual)
- We often write $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for the fitted value and $\hat{\varepsilon}_i$ for the residual.

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum \hat{\varepsilon}^2$$

---

example: show $\hat{\beta}_1$ is unbiased

- $E[\hat{\beta}_1] - \beta_1 = E\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right]$ (x is constant, so just need $E[Y]$)
- $E[Y_i] = \beta_0 + \beta_1 x_i$
- $E[\bar{Y}] = E\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n} E[\sum Y_i]$

$$= \frac{1}{n}\sum[\beta_0 + \beta_1 x_i] = \beta_0 + \beta_1 \frac{1}{n}\sum x_i$$

$$= \beta_0 + \beta_1 \bar{X}$$

$$\Rightarrow E[\hat{\beta}_1] = \frac{1}{\sum(x_i - \bar{x})^2} \cdot \sum E[(x_i - \bar{x})(Y_i - \bar{Y})]$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \sum (x_i - \bar{x})[E[Y_i] - E[\bar{Y}]]$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \sum (x_i - \bar{x})[\beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x})]$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \sum \beta_1 (x_i - \bar{x})^2$$

$$= \beta_1$$

Monday, February 23

exercise: find $\mathrm{Var}(\hat{\beta}_1)$

- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
- $\mathrm{Var}(\hat{\beta}_1) = \left[\sum(x_i - \bar{x})^2\right]^{-2} \cdot \sum(x_i - \bar{x})^2 \cdot \mathrm{Var}(y_i - \bar{y})$ (no independence $\neq \mathrm{Var}(Y_i) + \mathrm{Var}(\bar{Y})$)

claim:

- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})\, y_i}{\sum(x_i - \bar{x})^2}$

$$\Leftrightarrow \frac{\sum(x_i - \bar{x})\,\bar{y}}{\sum(x_i - \bar{x})^2} = 0$$

$$\Leftrightarrow \sum(x_i - \bar{x})^2 \bar{y} = 0$$

$$\Leftrightarrow \bar{y}\underbrace{\sum(x_i - \bar{x})^2}_{0} = 0$$

continuing: find $\mathrm{Var}(\hat{\beta}_1)$

- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\, y_i}{\sum(x_i - \bar{x})^2}$

$$\Rightarrow \mathrm{Var}(\hat{\beta}_1) = \frac{1}{\left[\sum(x_i - \bar{x})^2\right]^2} \cdot \left[\sum(x_i - \bar{x})^2\right] \cdot \mathrm{Var}(Y_i)$$

- $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$\Rightarrow \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\left[\sum(x_i - \bar{x})^2\right]}$$

$$\Rightarrow \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

(because of the linearity of $N(\mu, \sigma^2)$)

- We can now do inference on β_1 (always!)

---

recall:

- $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\varepsilon}_i^2$

$$= \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$ estimates σ^2

- this isn't unbiased.
- we'll use $\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$ which is unbiased.

$$\| $$

$$\frac{1}{n-2}\sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

---

definition: residual standard error, estimates $\hat{\sigma}^2$

- $RSE = \sqrt{\frac{1}{n-2}\sum \varepsilon_i^2} = \sqrt{\frac{SSE}{(n-2)}}$

---

fact:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

(MLE; just cancels $\frac{1}{n}$ from $\hat{\sigma}^2$)

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

recall:

- if

$$Z \sim N(0,1), \quad U \sim \chi^2(r)$$

then,

$$Z/\sqrt{U/r} \sim t(r)$$

- let $T = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/\sum(x_i-\bar{x})^2}} \sim N(0,1)$

$$\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}/(n-2)} \sim \chi^2(n-2)$$

$$= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{n\hat{\sigma}^2}{(n-2)\sum(x_i-\bar{x})^2}}}$$

(standard error of $\hat{\beta}_1$)

Wednesday, February 26

- $\bar{y} \sim N(\beta_0 + \beta_1\bar{x}, \sigma^2/n)$
- $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/\sum(x_i-\bar{x})^2\right)$
- $\hat{\beta}_0 \sim N\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right]\sigma^2\right)$
- $n\hat{\sigma}^2/\sigma^2 \sim \chi(n-2)$
- $T = Z/\sqrt{U/r} \sim t(r)$ with $U \sim \chi^2(r)$, $Z \sim N(0,1)$

$$= \frac{\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)\sigma^2}}}{\sqrt{n\hat{\sigma}^2/\sigma^2(n-2)}} = \frac{\hat{\beta}_0 - \beta_0}{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)^{1/2}} \cdot \frac{1}{\sqrt{\frac{\sigma^2}{(n-2)\sigma^2}}}$$

$$= \frac{\hat{\beta}_0 - \beta_0}{\underbrace{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)\frac{n\hat{\sigma}^2}{n-2}}}_{SE\,\hat{\beta}_0}} \sim t(n-2)$$

---

example:

- let's do a level-c confidence interval for β_0 (β_1 is similar).
- we want

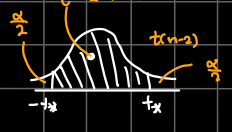
$$c = P(a(\vec{x}) < \beta_0 < b(\vec{x}))$$

$$= P\left(-t_{\alpha^*} < \frac{\hat{\beta}_0 - \beta_0}{SE\,\hat{\beta}_0} < t_{\alpha^*}\right)$$

$\Rightarrow t_{\alpha^*} = -q_T(\alpha^*/2, n-2)$

$\Rightarrow c = P(-t_{\alpha^*} SE_{\hat{\beta}_0} < \hat{\beta}_0 - \beta_0 < t_{\alpha^*} SE_{\hat{\beta}_0})$

$= P(\hat{\beta}_0 - t_{\alpha^*} SE_{\hat{\beta}_0} < \beta_0 < \hat{\beta}_0 + t_{\alpha^*} SE_{\hat{\beta}_0})$



---

fact:

- consider the distribution of $\hat{y}|x$.
- we know, $\hat{y}|x = (\beta_0 + \beta_1 x)$
- then,

$$\hat{y}|x \sim N\left(\beta_0 + \beta_1 x, \left(\frac{1}{n} + \frac{x-\bar{x}}{\sum(x-\bar{x})^2}\right)\sigma^2\right)$$

- which we can use to compute $SE\,\hat{y}|x$.

---

topic: multiple regression

- we consider models of the form

$$\begin{cases} \mu(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 & \text{(systematic component)} \\ y_i \sim N(\mu(Y|x), \sigma^2 & \text{(random component)} \end{cases}$$

- **question:** what does inference look like now?
  - we can do things like

$$H_0: \beta_2 = 0$$

  in this model.
  - we're asking what is the probability of data like what was observed if in fact the data comes from a system with component of

$$\mu(Y|x) = \beta_0 + \beta_1 x$$

- **question:** is the overall model significant?

(a p-value vs $\beta_0 = 0$ and $\beta_1 = 0$)

- **question:** more generally, our a model with systematic component

$$\mu(Y|x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

how does this compare with (is more better?)

$$\mu(Y|x) = \beta_0 + \sum_{i=1}^{p-r} \beta_i x_i, \quad 0 < r < p.$$

Friday, February 27

fact: sum of squares decomposition

- $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$

(y_i: observed y-value; $\bar{y}$: overall average; $\hat{y}_i$: model vs mean; residuals)

- $SS_{tot} = SS_{model} + SS_{residuals}$ [true for MLE]

(tot: total sum of squares)

---

pf/

- LHS $= \sum_i (y_i - \bar{y})^2$

$$= \sum[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum[(y_i - \hat{y}_i)^2 + \underbrace{2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\text{should be zero}} + (\hat{y}_i - \bar{y})^2]$$

- consider $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$. Recall, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$= \sum(y_i - \hat{y}_i)[\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})]$$

$$= \sum(y_i - \hat{y}_i)[\hat{\beta}_1(x_i - \bar{x})]$$

$$= \underbrace{\hat{\beta}_1 \sum(y_i - \hat{y}_i)x_i}_{=0 \text{ (MLE)}} - \underbrace{\hat{\beta}_1 \sum(y_i - \hat{y}_i)\bar{x}}_{=0 \text{ (MLE)}}$$

(the process of obtaining the MLE involves setting the derivative to 0)

- $SS_{total} = SS_{model} + SS_{residuals}$
- Suppose we want to test (systematic component)

  $H_0: \beta_1 = \dots = \beta_p = 0$ for $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$
- When H_0 is true, most of SS_{total} comes from $SS_{residuals}$.

  When H_0 is false, this won't generally be the case.
- We'll use this test statistic:

$$F = \frac{SS_{model}/(p-a)}{SS_{residuals}/(n-p-1)}$$

---

fact:

- $\frac{SS_{model}/(p-a)}{\sigma^2} \sim \chi(p-a)$
- $\frac{SS_{residuals}/(n-p-1)}{\sigma^2} \sim \chi(n-p-1)$

---

pf/

- Suppose $u_1 \sim \chi^2(r_1)$, $u_2 \sim \chi(r_2)$
- Then
  - $F = \frac{u_1/r_1}{u_2/r_2}$ ← used for comparing variances

  has an F-distribution with r_1 and r_2 degrees of freedom.

Monday, March 2

topic: Anova F-test

- Systematic:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

- Observations:

$$(x_1, \dots, x_p, y)$$

- More specifically

$$(x_{1i}, \dots, x_{pi}, y_i)$$

for the i^{th} observation.

- $H_0: \beta_0 = \dots = \beta_p = 0$

  the 'true' model is $y = \beta_1$ (i.e. $y = \bar{y}$)
- $F = \frac{SS_{model}/p}{SS_{resid}/(n-p-1)} \sim F(p, n-p-1)$

  ↳ doesn't include β_0

  ↳ p+1 is the number of coefficients in the model
- When H_0 is false, this tends to be large since SS_{model} is bigger when $\beta_0 \neq \dots \neq \beta_p$.
- We always compute a right-tailed probability.

  ↳ 1 - pf(x, p, n-p-1)

---

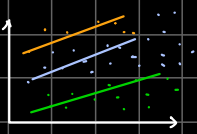
topic: including a categorical explainer

- To do this, we need a model of the form

$$y = \begin{cases} \beta_0 + \beta_1 x_1, & \text{setosa} \\ (\beta_0 + \beta_2) + \beta_1 x_1, & \text{versicolor} \\ (\beta_0 + \beta_3) + \beta_1 x_1, & \text{virginica} \end{cases}$$

vertical shift

x = sepal width



- For now, assume equal slopes
- Let $x_2 = \begin{cases} 1, & \text{versicolor} \\ 0, & \text{otherwise} \end{cases}$

  $x_3 = \begin{cases} 1, & \text{virginica} \\ 0, & \text{otherwise} \end{cases}$

$\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \; ; \; \varepsilon_i \sim N(0, \sigma^2)$

- $H_0: \beta_2 = \beta_3 = 0$
- In other classes, ANOVA is formulated in the context of a categorical variable explaining a quantitative one.

  ex/ do species of iris have different sepal lengths on average?

  $H_0: \mu_{set} = \mu_{virg} = \mu_{vers}$
- Take the model from our example, but excluding β_1 (sepal width):

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

$H_0: \beta_2 = \beta_3 = 0 \Rightarrow$ all groups means are equal

---

Wednesday, March 4

recall: the linear model

- $\begin{cases} u(x) = \beta_0 + \beta_1 x & \text{(systematic)} \\ Y \sim N(u(x), \sigma^2) & \text{(random)} \end{cases}$
- today, we'll use a poisson random component.

  $\Rightarrow Y \sim pois(\lambda(x))$
- A linear systematic component isn't ideal either (at least in the minion example)
- We'll instead consider:

$$\ln[\lambda(x)] = \beta_0 + \beta_1 x$$

$$\Rightarrow \lambda(x) = \exp[\beta_0 + \beta_1 x]$$

- Consider $w \sim pois(\lambda)$

$$\Rightarrow f(w) = \frac{\lambda^w e^{-\lambda}}{w!}$$

$$\Rightarrow L(y_i, \lambda) = \frac{\prod \lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{e^{-\lambda} \lambda^{\sum y_i}}{\prod y_i!}$$

$$= \frac{e^{-\exp(\beta_0 + \beta_1 x)} \exp[\beta_0 + \beta_1 x] \sum y_i}{\prod y_i!}$$

- $\ell(y_i, \lambda) = -\sum \exp(\beta_0 + \beta_1 x_i) + \sum y_i (\beta_0 + \beta_1 x) - \ln(\prod y_i)$
- Sadly, the system

$$\frac{\partial}{\partial \beta_0} \ell = 0$$

$$\frac{\partial}{\partial \beta_1} \ell = 0$$

doesn't have a closed form.

---

topic: generalized linear models (GLMs)

- we predict a function of the mean of the response variable y using a linear combination of the explainers $x_1, \dots, x_p$.
- we'll use $p = 1$ here.

$$\begin{cases} g(\mu_i) = \beta_0 + \beta_1 x_i & \text{systematic} \\ y_i \sim \text{EDM}(\mu_i) & \text{random} \end{cases}$$

(invertible)

- $g(\mu)$ is called the link function. it should be invertible (recover μ) and it should differentiable. MLEs it should also fit the shape of the data (convenient). (consider domain)
- recall the exponential family of distributions which have pdfs that can be written in the form

$$f(x;\theta) = \exp[k(x)p(\theta) + S(x) + q(\theta)]$$
$$= a(\theta)\exp[k(x)p(\theta) + S(x)]$$

($e^{q(\theta)}$; $\sum k(x_i)$ is sufficient; natural parameter; normalization constant)

- $\frac{1}{a(\theta)} \cdot \int_S \exp[k(x)p(\theta) + S(x)]$

---

example: poisson

- $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \cdot \exp[x \log\lambda - \log(x!)]$

($\sum x_i$ is sufficient; $p(\lambda) = \log\lambda$ is the natural param.; integrates to e^{λ})

- poisson regression: $\begin{cases} \log(\lambda_i) = \beta_0 + \beta_1 x_i \\ y_i \sim \text{pois}(\lambda_i) \end{cases}$ (mmmm...)
- other links are possible. $\log(\lambda)$ is called the canonical link because it corresponds to the natural parameter, $p(\lambda)$

---

definition: exponential dispersion model (EDM)

- this allows for a second parameter in the distribution. (dispersion)

$$f(x;\theta,\phi) = a(\theta,\phi)\exp\left[\frac{k(x)p(\theta)}{\phi} + S(x)\right]$$