statistics v.s. probability

probability : know about underlying distribution (population) ⟶ make predictions about sample

examples: • E of a random variable : tells you what the average value of your samples should be

• LLN : if you take lots of samples, then the sample average is likely to be close to the true underlying population average

• CLT : sample average follows a normal distribution

Statistics : start with data sample ⟶ try to learn sth about the underlying population that generated that sample.

<u>examples</u>: • Suppose you roll a die 6 times and you get a 1. 4/6 times

Q : do you conclude that the die is biased.

What if you got a 1 4000/6000 times ? same Q

where do you draw the line ?

• suppose there are 2 drugs for a disease drug 1 cured 9/10 people, drug 2 cured 9/10

Q: Is drug 2 better ?

What if drug 1 cured 9000/10,000, drug 2 cured 9000/10,000 ?

In other words : How do you decide if sth you observed is real or just caused by normal random fluctuations ?

The answers to these questions are necessarily a bit subjective, but the goal of this class is to be more quantitative about it.

• The main tool that we use for this is probability.

<u>conditional probability</u> "probability of B given A"

A, B are two events. $\mathbb{P}(B|A)$ = the prob. of B happening if we assume that A happens

examples: $\mathbb{P}(\text{roll a 2} \mid \text{roll an even \#}) = \frac{1}{3}$

$\mathbb{P}(\text{even \#} \mid 2) = 1$

Sneak peek: how to answer the Qs from above?

Calculate something like $\mathbb{P}\left(\begin{matrix}\text{getting a 1} \\ 4000/6000 \text{ times}\end{matrix} \mid \text{die is fair}\right)$

if this probability is tiny, then maybe you conclude that your assumption that the die is fair is incorrect

Bayes Rule

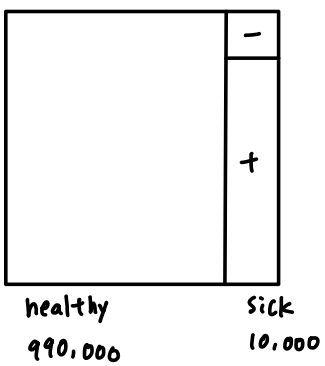
formula: $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)}$ (gross)

example:

Suppose there's a disease and 1% of people are sick. Also, the test for the disease is 99% accurate. Suppose you get a + test what is the probability that you are actually sick?

know: $\mathbb{P}(+ \mid \text{sick}) = 0.99$

want to know: $\mathbb{P}(\text{sick} \mid +)$



1,000,000 people

$$\mathbb{P}(\text{sick} \mid +) = \frac{\#\text{ sick + people}}{\text{total \# + ppl}}$$

$$= \frac{9900}{9900 + 9900}$$

$$= \frac{1}{2}$$

Order Statistics

X = some probability distribution (continuous)

take a bunch of iid samples $X_1, X_2, \dots X_n \sim X$
↳ independent and identically distributed

put the samples in order $Y_1 \le Y_2 \le Y_3 \dots \le Y_n$

text book # 6.3.3

$X = Exp(\theta = 3) = Exp(\lambda = \frac{1}{3})$ ← exponential distribution formula

→ pdf: $f(t) = \frac{1}{3}\exp(-\frac{1}{3}t)$
cdf: $F(t) = 1 - \exp(-\frac{1}{3}t) = P(X \le t)$

take 5 iid samples ~ ordered $Y_1 < Y_2 < \dots Y_5$

(a) find pdf of Y_3

(b) find $P(Y_4 < 5)$

(c) find $P(Y_1 > 1)$

goal: translate the condition "$Y_1 > 1$" into a condition involving the original samples.

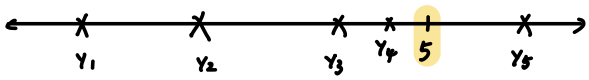
In this case: "$Y_1 > 1$" is equivalent to "all 5 samples are > 1"

$P(Y_1 > 1) = P(X_1 > 1 \text{ and } X_2 > 1 \text{ and } \dots \text{ and } X_5 > 1)$

$= P(X_1 > 1)\, P(X_2 > 1) \dots P(X_5 > 1)$

$= (1 - F(1))^5 = \exp(-\frac{1}{3})^5 \approx 0.19$

(b) $P(Y_4 < 5) = P(\text{something about the samples})$



P (At least 4 of the 5 samples are < 5)

$= P(4 \text{ samples below} \mid \text{sample above}) + P(\text{all 5 samples below})$

$= \binom{5}{4} F(5)^4 (1 - F(5))^1 + F(5)^5$

$= 5\,(1 - \exp(-\frac{5}{3}))^4 (\exp(-\frac{5}{3})) + (1 - \exp(-\frac{5}{3}))^5 \approx 0.96$

$P(X_1, X_3, X_4, X_5 < 5$ and $X_2 > 5) = F(5)^4 (1 - F(5))^1$

$\binom{5}{4}$ ways to rearrange

(a) find pdf of Y_3

find CDF first: CDF $F_{Y_3}(u) = P(Y_3 \le u)$

$P(Y_3 \le u) = P($ at least 3 samples are $\le u)$

$= P(3$ below, 2 above$) + P(4$ below, 1 above$) + P(5$ below$)$

$= \binom{5}{3} F(u)^3 (1-F(u))^2 + \binom{5}{4} F(u)^4 (1-F(u))^1 + F(u)^5$

$= 10\left(1-\exp\left(-\frac{u}{5}\right)^3\right) \exp\left(\frac{-u}{5}\right)^2 + 5\left(1-\exp\left(\frac{-u}{5}\right)\right)^4 \exp\left(\frac{-u}{5}\right) + \left(1-\exp\left(\frac{-u}{5}\right)\right)^5$

→ pdf: $\frac{d}{du}$ of this

6.3.5

X = some distribution with $\pi_{0.7} = 27.3$ (meaning $P(X \le 27.3) = 0.7$)

8 iid samples from X ⇝ order $Y_1 < Y_2 < Y_3 \cdots < Y_8$

find $P(Y_5 < 27.3 < Y_8)$

Y_5: at least 5 samples are below

Y_8: at most 7 samples are below

$\Rightarrow P(Y_5 < 27.3 < Y_8) = P($# of samples that are < 27.3 is 5, 6 or 7$)$

$= P(5$ below, 3 above$) + P(6$ below, 2 above$) + P(>$ below, 1 above$)$

$= \binom{8}{5}(0.7)^5(0.3)^3 + \binom{8}{6}(0.7)^6(0.3)^2 + \binom{8}{7}(0.7)^7(0.3)^1$

4/14 Maximum Likelihood Estimators (MLE)

- sampling coming from known family of distribution (Exp. Geom, etc.)
- goal estimate the parameter(s) based on the sample

What is a "good" estimate?

idea of MLE:

- Likelihood Function L (parameters) = $\mathbb{P}$ (getting the sample that we got)
- MLE = choice of parameters that maximizes this likelihood

Simple Example

You have a coin with $\mathbb{P}(\text{heads}) = p$

flip 10 times ~ THHTTTTHTH

how to guess p?

intuitive guess: $p = \frac{4}{10}$

Let's calculate the MLE

$$L(p) = \mathbb{P}(\text{THHTTTTHTH})$$

$$= (1-p)\cdot p \cdot p\,(1-p)^4 \cdot p\,(1-p)\cdot p$$

$$= p^4(1-p)^6$$

To find the MLE: do calculus to find the maximum

trick to make calculation easier:

consider $\ell(p) = \log L(p) = \log(p^4(1-p)^6)$

$$= 4\cdot\log p + 6\cdot\log(1-p)$$

set $0 = \frac{d\ell}{dp} = \frac{4}{p} - \frac{6}{1-p}$

Solve for p ⟶ $\boxed{p = \frac{4}{10}}$

More complicated example

pdf $f(t) = \lambda \cdot \exp(-\lambda t)$

Sample from $Exp(\lambda) \to x_1, x_2, \dots, x_n$

Find MLE for λ

$L(\lambda) = \mathbb{P}$ (getting the sample $x_1, x_2, \dots, x_n$)

$= f(x_1) \cdot f(x_2) \cdots f(x_n)$

(exp is a continuous distribution, so $\mathbb{P}$ of any specific sample is 0)

$L(\lambda) = \lambda \exp(-\lambda x_1) \cdot \lambda \exp(-\lambda x_2) \cdots \lambda \exp(-\lambda x_n)$

$= \lambda^n \cdot \exp(-\lambda x_1 - \lambda x_2 - \dots - \lambda x_n)$

$= \lambda^n \exp(-\lambda (x_1 + x_2 + \dots + x_n))$

log likelihood $\ell(\lambda) = \log L(\lambda) = \log(\lambda^n \cdot \exp(-\lambda(x_1 + \dots + x_n)))$

$= \log(\lambda^n) - \lambda(x_1 + \dots + x_n)$

$= n \cdot \log \lambda - \lambda(x_1 + \dots + x_n)$

Set $0 = \frac{d\ell}{d\lambda} = \frac{n}{\lambda} - (x_1 + \dots + x_n)$

solve $\to$ $\lambda = \frac{n}{x_1 + \dots x_n} = \frac{1}{\bar{x}}$ ⇢ Recall: $\mathbb{E}[\exp(\lambda)] = \frac{1}{\lambda}$

Hard example

Sample $x_1, x_2, \dots, x_n$ from $Unif(0, T) \to$ pdf $f(t) = \begin{cases} \frac{1}{T} & \text{for } 0 < t < T \\ 0 & \text{else} \end{cases}$

find MLE for T

✗ $L(T) = f(x_1) f(x_2) \cdots f(x_n)$

$= \frac{1}{T} \cdot \frac{1}{T} \cdot \dots \cdot \frac{1}{T}$

$= \frac{1}{T^n}$ ✗

"maximum" is ∞ at $T = 0$ (doesn't make sense Unif(0,0))

$L(T) = \left\{ \frac{1}{T} \text{ if } 0 < t < T, 0 \text{ else} \right\} \cdots \left\{ \frac{1}{T} \text{ if } 0 < x_n < T, 0 \text{ else} \right\}$

→ 只要有任何一個 x_i 不在 $0 < t < T$ 的 range, the whole thing is 0. (相乘)

$= \begin{cases} \frac{1}{T^n} & \text{if } 0 < \text{all of the } x_j < T \\ 0 & \text{if any } x_j > T \end{cases}$

0 can never be the maximum, so find the MLE we just need to find the maximum of $L(T) = \frac{1}{T^n}$ subject to the constraint that $T >$ all of the x_j

try calculus: $L'(T) = (-n)T^{-n-1} = 0$ no solutions because $L'(T) < 0$ for all T

$\frac{1}{T^n}$ is a decreasing function of T, so the maximum happens at the smallest possible value of T.

with constraint: $T = \max(x_1, x_2, \cdots, x_n)$

4/21

Method of moments & Percentile matching

method of moments (MoM):

- if X is a RV, then the k^{th} moment is $\mathbb{E}(X^k)$
- if you have a sample $x_1, \cdots, x_n$, you can calculate the k^{th} sample moment

$m_k = \frac{1}{n}(x_1^k + x_2^k + \cdots + x_n^k)$

- our distribution has some unknown parameter(s) θ
- MoM for estimating θ:
  - calculate theoretical moments (function of θ)
  - compare to sample moments
  - solve for θ

examples

6.4.10

$f(k) = (1-p)^{k-1} \cdot p$ $k = 1, 2, 3, \cdots$

Sample $x_1, \cdots, x_n \sim Geom(p)$

estimate p using MoM

theoretical | sample

① $E[\text{Geom}(p)] = \frac{1}{p}$ ⟷ (match) $m_1 = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}$

so our estimator for p is $\frac{1}{p} = \bar{x} \rightarrow \boxed{p = \frac{1}{\bar{x}}}$

6.4.15

pdf $f(x) = \frac{x^{\alpha-1} \cdot \exp(-\frac{x}{\theta})}{\Gamma(\alpha) \cdot \theta^{\alpha}}$

Sample $x_1, \ldots, x_n \sim \text{Gamma}(\alpha, \theta)$

estimate α and θ using MoM

→ could calculate $E(X)$ and $E(X^2)$ from pdf

given: $E = \alpha\theta$ and $\text{var} = \alpha\theta^2$ (var = $E(X^2) - E(X)^2$)

$\int_0^{\infty} x \cdot f(x)\, dx = \ldots$

theoretical | sample

① $E(X) = \alpha\theta$ | $m_1 = \frac{1}{n}(x_1 + \cdots + x_n)$

② $E(X^2) = \text{var}(X) + E(X)^2$ | $m_2 = \frac{1}{n}(x_1^2 + \cdots + x_n^2)$

$= \alpha\theta^2 + \alpha^2\theta^2$ | $m_3 = \frac{1}{n}(x_1^3 + \cdots + x_n^3)$

⋮

(match) → $\alpha\theta = m_1$, $\alpha\theta^2 + \alpha^2\theta^2 = m_2$ → solve for $\alpha\theta$ in terms of m_1, m_2

Percentile matching:

- to estimate unknown parameters, match theoretical percentile values with sample percentile
- # of different percentile points you read to compare = # of unknown parameters you need to estimate

6.4.21

Exp(θ) → pdf $f(x) = \frac{1}{\theta}\exp(-\frac{x}{\theta})$, $x > 0$

estimate θ by matching 0.5 percentile

We need to match the theoretical median to the sample median

- sample median let's call $\tilde{\pi}_{0.5}$
- Theoretical median $\pi_{0.5}$ is defined by $P(\text{Exp}(\theta) < \pi_{0.5}) = \frac{1}{2}$

$$\frac{1}{2} = \int_0^{\pi_{0.5}} f(x)\,dx = \int_0^{\pi_{0.5}} \frac{1}{\theta}\exp(-\frac{x}{\theta})\,dx = [-\exp(-\frac{x}{\theta})]_0^{\pi_{0.5}} \cdots = 1 - \exp(-\frac{\pi}{\theta})$$

$$\longrightarrow \quad \pi_{0.5} = \theta \cdot \log(2)$$

match $\pi_{0.5} = \tilde{\pi}_{0.5}$

$$\theta \log 2 = \tilde{\pi}_{0.5}$$

→ our estimate is $\theta = \frac{\tilde{\pi}_{0.5}}{\log(2)}$

Other example

distribution: Unif (a, b)

let's use 25^{th} and 75^{th} percentile

sample percentiles $\tilde{\pi}_{\frac{1}{4}}$ and $\tilde{\pi}_{\frac{3}{4}}$ — match

theoretical percentiles $\pi_{\frac{1}{4}} = a + \frac{1}{4}(b-a)$ — match

$\pi_{\frac{3}{4}} = a + \frac{3}{4}(b-a)$ — match

solve for a, b in terms of $\tilde{\pi}_{\frac{1}{4}}$ and $\tilde{\pi}_{\frac{3}{4}}$

4/28 Interval estimation

- So far: all point estimators (guessing a single # for the unknown parameter)
- more useful to report a range of values along with some level of confidence that your range captures the true value.

Simple example

Sample $X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ (μ: unknown; σ^2: assume to be known)

Find a 95% confidence interval for μ

- meaning: we will give a range, expressed as a function of the samples $X_1, \dots, X_n$, such that μ is within the range with 95% probability.
- how do we do this?

idea: something to do with $\bar{x}$

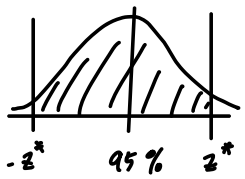
each $X_i \sim N(\mu, \sigma^2)$

so $X_1 + \dots + X_n \sim N(n\mu, n\sigma^2)$

so $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

one final transformation: $\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

let's write down an interval that has 95% probability:



there is some number z^* so that

$$\mathbb{P}\left(-z^* < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z^*\right) = 0.95$$

↓ algebraic manipulation

$$\mathbb{P}\left(\bar{x} - z^*\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^*\frac{\sigma}{\sqrt{n}}\right)$$

conclusion: We can say that the range

$\boxed{\bar{x} \pm z^*\frac{\sigma}{\sqrt{n}}}$ has a 95% chance of containing μ

($z^*\frac{\sigma}{\sqrt{n}}$: margin of error)

- How to find the number z^*?

use the lookup table for 95%: $z^* = 1.96$

Q: what if the X_i are not normal?

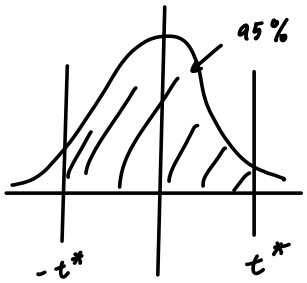
use central limit theorem: it doesn't matter if the X_i are normal or not, $\bar{x}$ is always approximately normal if n is big

Q: what do we do if σ is also unknown?

We can try just replacing σ by S (the sample st. dev) ↓ divided by n-1 instead of n everywhere it appears.

problem. $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ does not follow $N(0,1)$

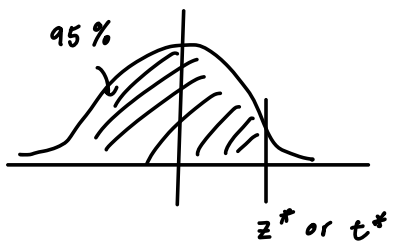
- It follows something called a T distribution with n-1 degrees of freedom.



we just have to use a different look up table

Q: what if we want μ in the range $[\bar{x} - \_\_, \infty)$?

instead of range $\bar{x} \pm \_\_\_$



5/5 C.I. for difference of 2 means

simplest setting: $X \sim N(\mu_x, \sigma_x^2 = \text{known})$, $Y \sim N(\mu_y, \sigma_y^2 = \text{known})$ } independent

goal: CI for $\mu_x - \mu_y$

We know that $\bar{x} \sim N(\mu_x, \frac{\sigma_x^2}{n_x})$, $\bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n_y})$ } still independent

$\Rightarrow \bar{x} - \bar{Y} \sim N(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y})$

$\Rightarrow \frac{\bar{x} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$

$\rightsquigarrow$ CI for $\mu_x - \mu_y$ is $(\bar{x} - \bar{Y}) \pm z^* \cdot \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$

↖ finding z^* is exactly the same as before

What happens if:

- σ_x, σ_y are unknown?
- X, Y are not normal?
- X, Y are not independent?

Midterm solution

2 $X \sim N(\mu, \sigma^2)$

estimate μ, σ^2 using percentile matching w/ $p = 0.3, 0.7$

how to find $\pi_{0.3}$ $\pi_{0.7}$ of $N(\mu, \sigma^2)$?

- transform from $N(0,1)$: if we know that 30^{th} % of $N(0,1)$ is M,

then $\pi_{0.3} = \sigma \cdot M + \mu$?

use z table: for $N(0,1)$, 30^{th} % $= -0.525$ $\rightsquigarrow$ $\pi_{0.3} = -0.525\sigma + \mu$

$70^{th} = 0.525$ $\pi_{0.7} = 0.525\sigma + \mu$

$\rightarrow$ $\hat{\mu} = 0$ $\hat{\sigma} = 1.924$

using MoM:

$E(X) = \mu$ $\qquad m_1 = \frac{1}{n}\sum x_i$

$E(X^2) = \text{var}(X) + E(X)^2$ $\qquad m_2 = \frac{1}{n}\sum x_i^2$

$= \sigma^2 + \mu^2$

4

$X \sim Poi(\theta)$

Prior $h(\theta) = \frac{5}{\theta^6}$ $\quad \theta > 1$

Bayes Rule: posterior $h(\theta \mid x_1, \dots, x_n) = \frac{g(x_1, \dots, x_n; \theta) \cdot h(\theta)}{C}$

$= \frac{(e^{\theta} \cdot \frac{\theta^{x_1}}{x_1!}) \cdots (e^{-\theta} \frac{\theta^{x_n}}{x_n!}) \cdot \frac{5}{\theta^6}}{C}$

$= C \cdot e^{-n\theta} \cdot \theta^{x_1 + \dots + x_n - 6}$

$= Gam(\alpha = x_1 + \dots + x_n - 5, \ \beta = \frac{1}{n})$

Gam (α, β):

pdf $f(x) = C \cdot x^{\alpha-1} \cdot e^{-\frac{x}{\beta}}$

$x > 0$

May 12

confidence intervals for percentiles

$X \sim$ some continuous distribution

goal: find CI for the median $m = \pi_{0.5}$ (or some other percentile)

- start with the point estimate (sample percentile)

example

$n = 9$ samples, want a CI for the median m

point estimate $= Y_5$

possible intervals: (Y_4, Y_6)

(Y_3, Y_7)

(Y_2, Y_8)

隨機選都可用

What is the confidence level of these intervals?

(Y_3, Y_7): $\mathbb{P}(Y_3 < m < Y_7) = \mathbb{P}$(the # of samples that are $\le m$ is 3, 4, 5, 6)

$$= \binom{9}{3} \cdot \mathbb{P}(X<m)^3 \cdot \mathbb{P}(X>m)^6 + \cdots \binom{9}{6} \mathbb{P}(X<m)^6 \mathbb{P}(X>m)^3$$

$$= \binom{9}{3}\left(\tfrac{1}{2}\right)^3\left(\tfrac{1}{2}\right)^6 + \cdots + \binom{9}{6}\left(\tfrac{1}{2}\right)^6\left(\tfrac{1}{2}\right)^3 = \ldots$$

example

$n = 12$, fin CI for $\pi_{0.3}$

- point estimate: $(12+1)(0.3) = 3.9 \rightsquigarrow \tilde{\pi}_{0.3} = (0.1)\, Y_3 + (0.9)\, Y_4$
- let's just pick Y_4 to serve as our center
- options for intervals: (Y_3, Y_5) ←

(Y_2, Y_6)

⋮

$$\mathbb{P}(Y_3 < \pi_{0.3} < Y_5) = \mathbb{P}(\#\text{ samples} < \pi_{0.3} \text{ is 3 or 4})$$

$$= \binom{12}{3} \cdot P(X < \pi_{0.3})^3 \mathbb{P}(X > \pi_{0.3})^9 + \dots$$

$$= \binom{12}{3} (0.3)^3 (0.7)^9 + \dots$$

example

n samples (n is big)

We want to find a 95% CI for the median m

- point estimate = $Y_{\frac{n}{2}}$
- only consider intervals of the form $(Y_{\frac{n}{2}-k}, Y_{\frac{n}{2}+k})$

Q: find the best k so that the confidence level of this interval is $\approx 95\%$

$$\mathbb{P}(Y_{\frac{n}{2}-k} < m < Y_{\frac{n}{2}+k}) = \mathbb{P}(\underbrace{\#\text{ samples} < m}_{W \sim \text{Bin}(n, \frac{1}{2})} \text{ is in the interval } (\tfrac{n}{2}-k, \tfrac{n}{2}+k))$$

$$= \mathbb{P}(\tfrac{n}{2}-k < \text{Bin}(n, \tfrac{1}{2}) < \tfrac{n}{2}+k)$$

normal approximation / CLT: when n is big, $\text{Bin}(n,p) \approx N(np, np(1-p))$

in our case: $W \approx N(\frac{n}{2}, \frac{n}{4})$

$(Y_{\frac{n}{2}-k}, Y_{\frac{n}{2}+k})$ has confidence $\approx \mathbb{P}(\frac{n}{2}-k < N(\frac{n}{2}, \frac{n}{4}) < \frac{n}{2}+k)$

$$= \mathbb{P}\left(\frac{-k}{\sqrt{\frac{n}{4}}} < N(0,1) < \frac{k}{\sqrt{\frac{n}{4}}}\right)$$

set = 95%, solve for k:

$$\frac{k}{\sqrt{\frac{n}{4}}} = 1.96$$

$$k = \frac{1.96}{2}\sqrt{n}$$

for example if $n = 100$, then $k = \frac{1.96}{2} = 9.8 \approx 10$

So an approximate 95% CI for m is (Y_{40}, Y_{60})

May 19

hypothesis testing

Q: are UCLA students smarter than average?

Assume: average IQ = 110

IQ of UCLA students $\sim N(\mu, \sigma^2 = 100)$

Q: is $\mu > 110$?

Suppose we sample $n = 16$ UCLA students and get $\bar{x} = 113.5$

- Is this enough evidence to conclude that $\mu > 110$?

set up:

null hypothesis $H_0: \mu = 110$

alternative hypothesis $H_1: \mu > 110$

We ask the question: if we assume H_0 is true, then what is the probability of getting a sample at least as extreme as the one we actually got?

assuming H_0: $X \sim N(\mu = 110, \sigma^2 = 100) \longrightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

(assuming)

calculate $\mathbb{P}(\bar{x} \geq 113.5) = \mathbb{P}\left(\frac{\bar{x} - 110}{\frac{10}{\sqrt{16}}} \geq \frac{3.5}{\frac{10}{\sqrt{16}}}\right) = \mathbb{P}(N(0,1) \geq 1.4) \approx 0.08$ → called the p-value

- Is this p-value small enough to conclude that H_0 is false?
- It depends on what significance level we are using for the test
- For example, $\alpha = 10\%$, then $8\% < 10\% \rightsquigarrow$ reject the H_0 in favor of H_1

  using $\alpha = 5\%$, then $8\% > 5\% \rightsquigarrow$ fail to reject the H_0

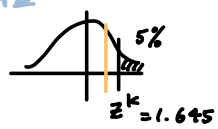
you will find the H_0 value for μ is outside the appropriate CI

⇕

the p-value of your sample is $< \alpha$

another way to think about it:

- when we get our sample $\bar{x}$, calculate the "test statistics" $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
- compare z to whatever critical threshold is (門檻)
- in this case, the critical threashold is



8.1.4

$X \sim N(\mu, \sigma^2 = \text{unknown})$

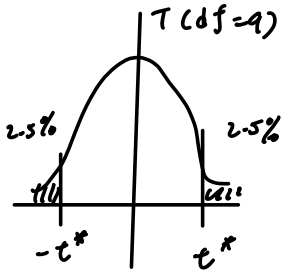
$H_0 : \mu = 7.5$

$H_1 : \mu \neq 7.5$

$n = 10$, $\bar{x} = 7.55$, $s_x = 0.1027$

do a hypothesis test at 5% significance

- calculate our test statistic $T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.55 - 7.5}{0.0127/\sqrt{10}} \approx 1.539$
- what is our critical threashold for this test?



$t^* = 2.26$

because our T is not extreme enough for the threshold of this test, we fail to reject H_0.

5/26

Test for variances

general structure of hypothesis tests:

1) get a point estimator to base the test on (for means: $\bar{x}$)

2) transform it into a "test statistics" that we know follows some special distribution (For means: $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ or $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$)

3) Compare your test statistics to the appropriate critical values.

Test for variance:

$X \sim N(\mu, \sigma^2)$

$n = 11$ samples

$H_0: \sigma^2 = 525$

$H_1: \sigma^2 < 525^2$

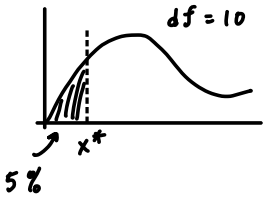
$\alpha = 5\%$

base estimator s^2

test statistic is called $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{10\,s^2}{525^2}$

- this follows a distribution called the "χ^2 distribution with $n-1=10$ d.o.f"

Find the critical region:



use the χ^2 lookup table: $x^* = 3.94$

critical region: reject H_0 if $\chi^2 < 3.94$

$s^2 < \frac{(3.94)(525)^2}{10}$

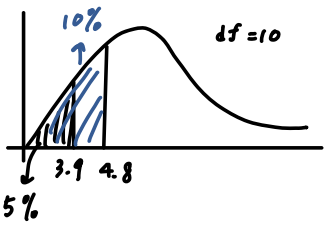
Let's say we take an actual sample and get $s^2 = 113, 108.5$

$\rightarrow \chi^2 = \frac{10(113, 106.5)}{525^2} = 4.1 > 3.9$, we fail to reject

---

What is the p-value of this sample?

$p\text{-val} = P(\chi^2(df=10) \leq 4.1) \in (5\%, 10\%)$



↳ 沒有辦法找到 exact value → approx.

$X \sim N(\mu_X, \sigma_X^2)$ $\quad Y \sim N(\mu_Y, \sigma_Y^2)$

$n_X = 11$ samples $\quad n_Y = 7$ samples

$H_0 : \sigma_X = \sigma_Y$

$H_1 : \sigma_X > \sigma_Y$ / $\sigma_X < \sigma_Y \Rightarrow$ 5% left tail

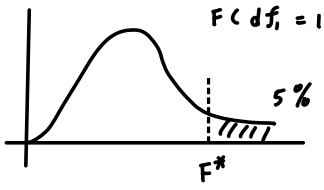
start with S_X^2, S_Y^2 $\quad$ ↳ flip $F = \frac{S_Y^2}{S_X^2}$

test statistic $= F = \frac{S_X^2}{S_Y^2}$

· this follows the F distribution with $df_1 = n_X - 1$,

$df_2 = n_Y - 1$

Find the critical region:



use the F look up table : $F^* = 4.06$

so our critical region for rejecting H_0 is $F = \frac{S_X^2}{S_Y^2} > 4.06$

For $\sigma_X < \sigma_Y$: instead define $F = \frac{S_Y^2}{S_X^2} \sim F(df_1 = n_Y - 1, df_2 = n_X - 1)$

and then do the same thing

6/2 ① Finals OH: Tues (June 7) 2-4 pm (in person MS 6139)

② Regrade grades ⇒ grade scope to grade book

Hypothesis Tests for Median

"wilcoxon signed rank test"

example (8.5.1)

We want to test

H_0: m = 40,000

H_1: m < 40,000

α = 5%

(n = 13)

sample: 41195 39485 41229 36840 38050 40890 38345

34930 39245 31031 40780 38050 30906

Step 1: look at the differences:

1195	-515	1229	-3160	-1950	890	-1655
(+5)	(-1)	(+6)	(-10)	(-8.5)	(+4)	(-7)

-5070	-755	-8969	780	-1950	-9040
(-11)	(-2)	(-12)	(+3)	(-8.5)	(-13)

(-8.5 appears twice)

Step 2: assign "signed ranks"

Step 3: calculate

W = sum of signed ranks = -1 -2 +3 +4 +5 +6 -7 -8.5 -8.5 -10 -11 -12 -13

= -55

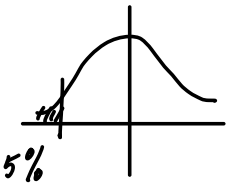
Step 4: apply the normalizing transformation:

$$Z = \frac{W}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \sim N(0,1)$$

why is this true? it's complicated (textbook proofs)

$$Z = \frac{-55}{\sqrt{\frac{(13)(14)(27)}{6}}} = -1.922$$

Step 5: compare to appropriate critical value



crit. region: Z < -1.645

our Z was -1.92, so reject