Last time:

Sec. 8 Normal curve
CLT

Today:
Sec 9 Inference. Hypotheses test.

prob. model: experiment / assumption → prob. / other stats. (expectation, sd, percentiles, ...)
Inference: how they are generated ←

Sec 9.1 Hypotheses test.

Step 1. Make Hypotheses.
1) Null hypothesis H_0: how you can simulate data, use this to calculate prob.s "$\mu = \mu_0$"
2) Alternative hypothesis H_a: may involve no chances
usually "$\mu \neq \mu_0$" "$\mu > \mu_0$"

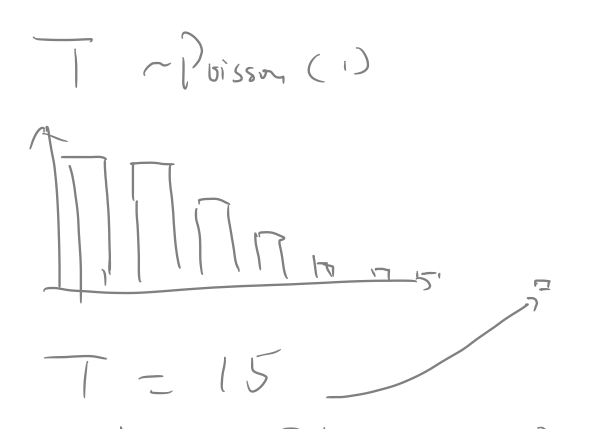
Step 2: Select a test statistic: which helps you make a decision
usually we prefer to select a statistic T such that larger value of T makes you lean towards alt. hypothesis H_a
or small value, but not both

Step 3. Find the distr. of the test statistic under null hypothesis H_0.

Step 4. Calculate the observed value of the test statistic and see if it looks consistent with the distr.

→ p-value: it is the chance that you see equally or even extremer data / test statistic under H_0
This is a prob., hence $\in [0, 1]$
"Small" value means you should reject H_0.
↳ up to you. usually we set a cutoff to 5% or 1%

e.g. $T \sim$ Poisson (1)



$T = 15$
p-value $= \mathbb{P}(T \geq 15)$

Example 1 Speed of Light.
In 1926, Michelson reported $c_0 = 299,796$ km/s
$c = 299,792.458$ km/s

Suppose 150 measurements, with mean c_0.
sd $\sigma = 50$ km/s (from 150 sample)

Are these data consistent w/. the model that they are i.i.d. with mean c? Or are they too large?

Hypotheses test.

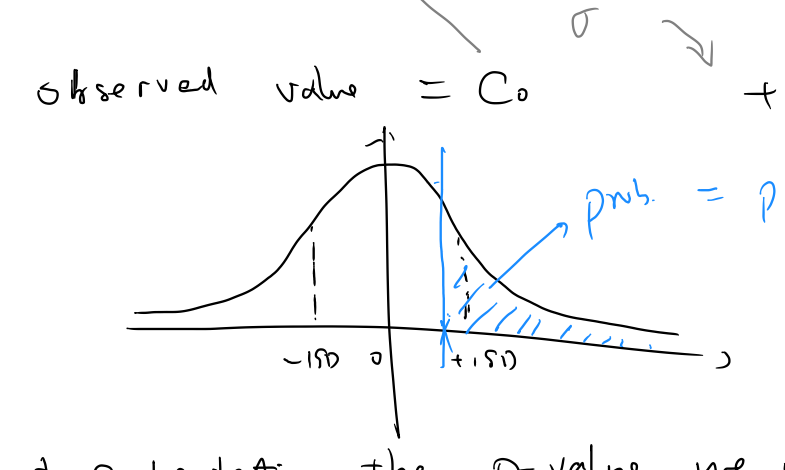
Step 1. H_0: model is good, i.e. $X_1, X_2, \dots, X_{150}$ i.i.d. w/. mean $\mu = c$
H_a: $\mu > c$
Note. we say nothing about sd(X)

Step 2. Select $\bar{X} := \frac{1}{150} \sum_{i=1}^{n} X_i$ (sample mean)

Step 3. Distri. of $\bar{X}$ under H_0
Since $n = 150$ is large, use CLT and take sample sd σ as an estimated population sd
i.e. $SD(X_i) = \sigma$, for any i
$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{150}} \approx 4$ (km/s)

Back to Step 2. probably we could use $\frac{\bar{X} - c}{SD(\bar{X})}$ as the test statistic $= \frac{c_0 - c}{\sigma}$

Step 4. observed value $= c_0 = +0.9$ SD



Without calculating the p-value, we may claim this as consistent.

p-value: recall normal cdf, this is $1 - \Phi(0.9) \approx 0.18 > 5\%$
Conclusion: Consistent.

Example 2. Emily Rosa's experiment.

"therapeutic touch" TT
A - performs TT → guess the hand (TT)
B - doing the test. → gives a hand

150 trials 70 times correct

You may do hypotheses test, but not necessary.
Since pure guess has an expectation of 75 correct guesses

H_0: $p = \frac{1}{2}$ Test statistic: Binom($n = 150$, $p = \frac{1}{2}$)
H_a: $p > \frac{1}{2}$

Sec 9.2 A/B-testing.

Example 3 Treatment for pain.

31, SRS → treatment 15 → 9 relief, 6 not
→ Control 16 → 2 relief, 14 not

What can we say about the treatment?

H_0: The treatment does nothing, i.e. the difference is due to random assignment
H_a: — — — — — something

Test statistic. $X :=$ the # of treated patients who have pain relief

		relief	not
A	treat	$X = 9$	6
B	control	2	14

Under H_0, no matter treat or not, 11 out of 31 patients will have pain relief.

$X \sim$ HyperGeom ($N = 31$, $G = 11$, $n = 15$) $\mathbb{E}X = \frac{nG}{N} \approx 5.32$

Since we are considering both tails, a better test statistic is $|X - \mathbb{E}X| = |X - 5.32|$

Observed value $= |9 - 5.32| = 3.68$

p-value: $\mathbb{P}(|X - 5.32| \geq 3.68)$
$= \mathbb{P}(X \geq 9 \text{ or } X \leq 1)$
< 0.01 — very significant! (evidence towards H_a)

Conclusion: effective!

Remark: In Data 8 we obtain the same conclusion by simulation:
① Pool the two groups
② randomly permute the pooled sample
③ See if it is ≥ 9 or ≤ 1
④ Repeat for many times and compute a empirical p-value
$\frac{\text{\# of trials} \geq 9 \text{ or } \leq 1}{\text{total \# of trials}}$

Sec. 9.3 Confidence Interval (CI)

What?
an interval of estimates of a fixed but unknown parameter based on data in a random sample

Why? (vs. point est.)
Point estimation could be different if we happen to use another set of data. So it does not take the variation into account.

Construct CI for an unknown population mean.

n (large) sample $X_1, X_2, \dots, X_n$ i.i.d.
population mean $= \mu$ (not random)

What we know: $X_1, X_2, \dots, X_n$
need: sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
(sample) sd: σ

CLT tells us $\frac{\bar{X} - \mathbb{E}\bar{X}}{sd(\bar{X})}$ follows standard normal curve

$\mathbb{E}\bar{X} = \mu$ $sd(\bar{X}) = \frac{1}{\sqrt{n}} \sigma$

$\mathbb{P}\left(-2 \leq \frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}}\sigma} \leq +2\right) \approx 95\%$

$\mathbb{P}(\mu \in (\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})) = 0.95$ (random)

$\mathbb{P}(X \in (3, 5))$ (random)

note: μ is not random. CI is random with 95% to cover μ.

Example 1 Commute distance

In a large city
SRS 600 workers sample mean $= 19$ miles
(sample) sd $= 13$ miles.

Find 95% CI for the average commute distance of all the workers in the city. → μ, the unknown fixed parameter.

600 is large, but also small in comparison to the total # of workers in the large city. treat as i.i.d. sample
sd of sample mean $= 13/\sqrt{600} \approx 0.53$

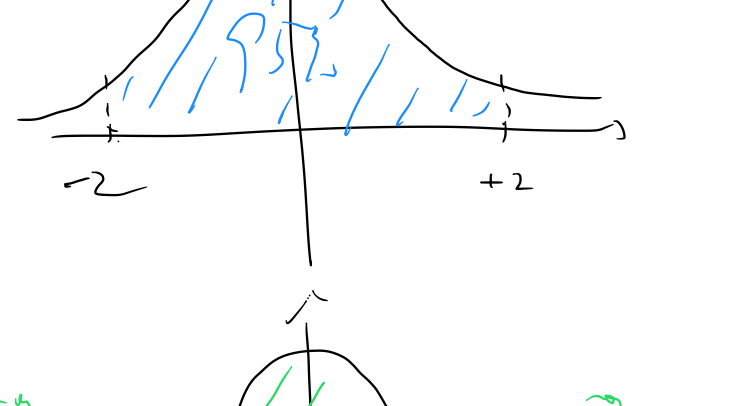
CI: $19 \pm 2 \times 0.53$.

Example 2 Undecided Voters

SRS of 400 voters in a state. 23% are undecided.
95% CI of percentage of undecided voters in the state

each individual $\sim$ Bernoulli (p) ← the parameter
Sample sd $= \sqrt{23\%(1 - 23\%)}$
sd of sample mean $= \sqrt{23\%(1-23\%)} / \sqrt{400}$
CI: $23\% \pm 2 \times$ ○

What if 99% CI instead of 95%?



for 95%, use $\pm(\Phi^{-1}(0.95))$.
for 99%, use $\pm(\Phi^{-1}(0.995))$

$2 \approx \Phi^{-1}(0.975)$