July 29, 2019

NOTE: If CLT says that (for large n) $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Then this means

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

A weakening of this result (when, for example, you don't know σ) is to replace σ by S, where S is given by

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$

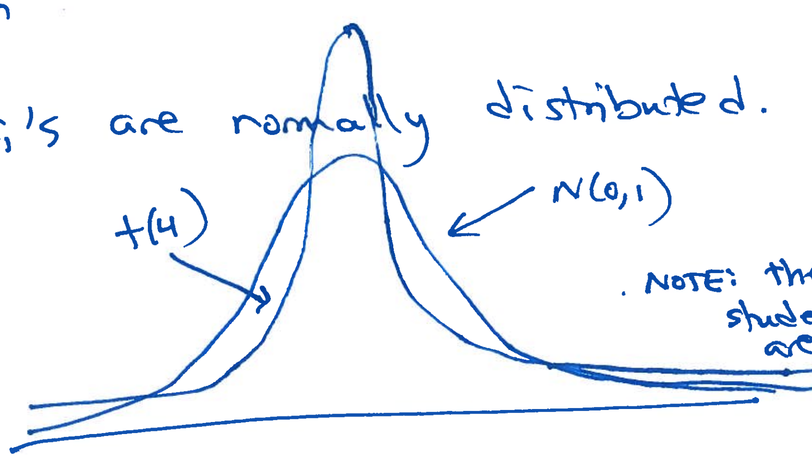
i.e., you obtain

degrees of freedom

$$T = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

student's t distribution

if the X_i's are normally distributed.



NOTE: the tails of a student's t distribution are a bit fatter than a Gaussian.

NOTE: It is a common result that as $n \to \infty$, $t(n-1) \to N(0,1)$.

MOTIVATION FOR CONFIDENCE INTERVALS:

All CI-like results come from some CLT-like result, or some knowledge about the distribution of a statistic.

EX: Let $X_1, \ldots, X_n$ be a random sample from a normal r.v. with unknown mean μ and known variance σ^2. Under the CLT,

$$\mathbb{P}\left(-z^*_{\frac{\alpha}{2}} \leq \underbrace{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}_{\text{standardized version of } \overline{X}_n} \leq z^*_{\frac{\alpha}{2}}\right) = 1-\alpha$$

exactly

quantile threshold selected to leave $\frac{\alpha}{2}$ probability in left tail

quantile you get from tables or things like qnorm

Let's take this expression and seek to isolate μ.

$$\Rightarrow \mathbb{P}\left(-\overline{X}_n - z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X}_n + z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\overline{X}_n + z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X}_n - z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\overline{X}_n - z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z^*_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

deterministic parameter

interval with random endpoints

INTERPRETATION: Let $a = \bar{X}_n - z^*_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ and $b = \bar{X}_n + z^*_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

#1: ~~"100(1−α)% of the data lie between a and b."~~

#2: ~~"100(1−α)% of the $\bar{X}$'s lie between a and b."~~

#3: "There is a 100(1−α)% chance that the true population mean μ lies between a and b."

"OBJECTION": Suggests that μ is stochastic and downplays notion that a and b are random endpoints.

CORRECTION? "The random interval [a, b] is constructed in such a way that it contains μ 100(1−α)% of the time."

#4: "The confidence interval over my particular random sample is [a, b]. Similarly-constructed intervals, computed over many different random samples, contain the true population mean μ with probability 1−α."

The prior confidence interval

$$\bar{X}_n \pm z^*_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is exact if $X_1, \dots, X_n$ are ~~normal~~ normal and σ is known.

Both assumptions are unrealistic.

WEAKENING #1: If the X_i are not normal — but are not severely non-normal — then the interval

$\bar{X}_n \pm z^*_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is approximate rather than exact.

(Justification: CLT.)

WEAKENING #2: If σ^2 is unknown and must be estimated with S^2, we end up with

$$\bar{X}_n \pm z^*_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

This substitution is justified by Slutsky's Theorem.

In general, you can make this approximation a little more exact by using $t^*_{\frac{\alpha}{2}, n-1}$ instead of $z^*_{\frac{\alpha}{2}}$.

If we replace $z^*_{\frac{\alpha}{2}}$ by $t^*_{\frac{\alpha}{2}, n-1}$ and use

$$\bar{X}_n \pm t^*_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

this $100(1-\alpha)\%$ CI is exact ~~with~~ when the X_i are normal (but σ unknown) and an approximation otherwise.

THE DEMOIVRE – LAPLACE THEOREM: (precursor for CLT when the X_i's are Bernoulli r.v.'s)

Let $X_i \sim Ber(p)$ be independent for $i = 1, \dots, n$. Call $X = X_1 + \dots + X_n$. (Aside: $X \sim Bin(n, p)$.)

Then

$$\mathbb{P}\left(a < \frac{(X_1 + \dots + X_n) - np}{\sqrt{np(1-p)}} < b\right) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

In other words

$$= \Phi(b) - \Phi(a)$$

$$X \sim Bin(n, p) \underset{\text{approximate}}{\sim} N(np, np(1-p))$$

$$\hat{p} = \hat{\pi} = \bar{X}_n = \frac{1}{n} X \sim \frac{1}{n} N(np, np(1-p)) = N\left(p, \frac{p(1-p)}{n}\right)$$

ASIDE: I will try to reserve π for the true population ~~parameter~~ proportion and $\hat{\pi}$ for the sample proportion.

$$\mathbb{P}\left(-z^*_{\frac{\alpha}{2}} \le \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \le z^*_{\frac{\alpha}{2}}\right) \approx 1 - \alpha \quad \text{(by D–L)}$$

$$\Rightarrow \mathbb{P}\left(-z^*_{\frac{\alpha}{2}} \le \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \le z^*_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

$$\Rightarrow \mathbb{P}\left(\hat{\pi} - z^*_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \le \pi \le \hat{\pi} + z^*_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \approx 1 - \alpha$$

These are the endpoints for our $100(1-\alpha)\%$ CI for π.

Qn: when is the approximation under DeMoivre-Laplace any good?

ANS: There is a heuristic that when $\min\{n\pi, n(1-\pi)\} \geq 10$, the approximation is fine.

If π is unavailable, just use $\hat{\pi}$.

Qn: How do I construct a CI for σ^2?

RESULT: If $X_1, \dots, X_n$ are i.i.d. normal r.v.'s with mean μ and variance σ^2, then

degrees of freedom

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

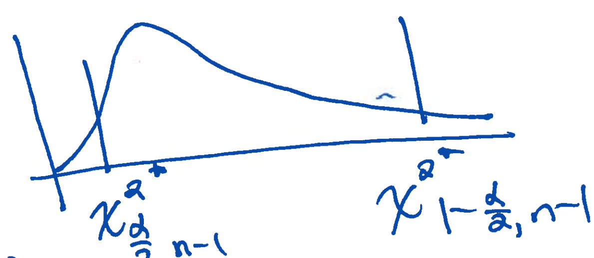
chi-squared

NOTE: $\chi^2(n)$ is, by definition, a r.v. generated by summing squared independent standard normal random variables.

EX: If $X \perp Y$ and $Z = X^2 + Y^2$ and $X \sim N(0,1)$ with $Y \sim N(0,1)$, then $Z \sim \chi^2(2)$.

Using this result,

$$\mathbb{P}\left(\chi^{2*}_{\frac{\alpha}{2}, n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^{2*}_{1-\frac{\alpha}{2}, n-1}\right) = 1 - \alpha$$



Gives $100(1-\alpha)\%$ CI formula for σ^2.

$$\Rightarrow \mathbb{P}\left(\frac{(n-1)S^2}{\chi^{2*}_{1-\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^{2*}_{\frac{\alpha}{2}, n-1}}\right) = 1 - \alpha$$

Qn: For a theorem — like the one we just used or student's theorem — how much abuse can the normality assumption take?

As long as the r.v./data satisfy the following conditions, most results that we see in this class that depend upon the normality assumption are robust to its violation.

1. No heavy tails (for a r.v.) or no outliers (if it is data drawn from a r.v.)
2. No skewness, particularly severe skewness.
3. No multi-modality, i.e., you distributions should be unimodal.

HYPOTHESIS TESTING:

EX: Suppose that $X_1, \ldots, X_n$ is drawn from a normal r.v. in an independent way. Suppose that you want to test $H_0: \sigma^2 = 25$ and $H_1: \sigma^2 \neq 25$.

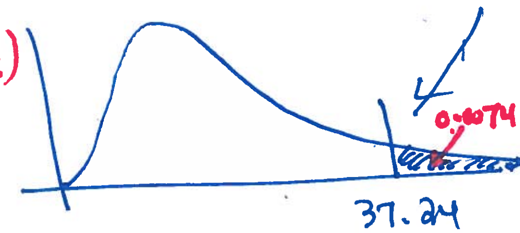
IDEA: If $X_i \sim N(\mu, \underbrace{\sigma^2 = 25}_{\text{variance under } H_0})$, then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

Suppose that $n=20$ and we find that $S^2 = 49$.

we calculate $\frac{(n-1)S^2}{\sigma^2} = \frac{(20-1)(49)}{25} \approx 37.24$.

1-pchisq(37.24, 19)

≈ 0.007404



this tail probability encodes how rare it is to see ≥ 37.24 if H_0 is true.