AP Statistics in 75 minutes

Data: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\underbrace{\mu_0, \sigma_0^2}_{\text{true but unknown population parameters}})$

Estimators:

sample average $\quad \hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

sample variance $\quad \widehat{\sigma^2_n} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$

Exact sampling distributions:

$$\hat{\mu}_n \sim N(\mu_0, \sigma_0^2/n)$$

$$\widehat{\sigma^2_n} \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}\frac{1}{\sigma_0^2}\right)$$

These random variables are independent.

Confidence interval: Exact $100 \times (1-\alpha)\%$ CI for the mean

$$\underbrace{\hat{\mu}_n}_{\text{point estimate}} \pm \underbrace{\underbrace{t^*_{n-1}\left(1-\frac{\alpha}{2}\right)}_{\text{"t score"}} \underbrace{\sqrt{\frac{\widehat{\sigma^2_n}}{n}}}_{\text{standard error}}}_{\text{margin of error}}$$

Hypothesis test:

$H_0: \mu_0 = h \quad \leftarrow$ point null

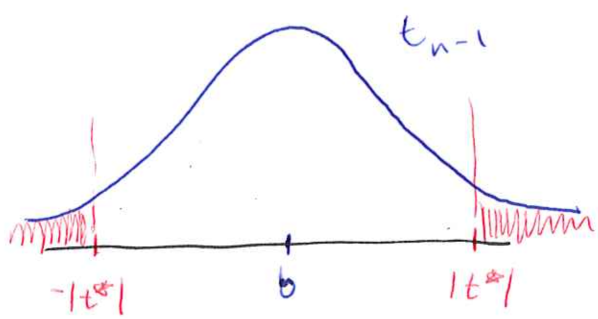
$H_A: \mu_0 \neq h \quad \leftarrow$ two-sided alternative

(test statistic) $\quad T_n = \frac{\hat{\mu}_n - h}{\sqrt{\frac{\widehat{\sigma^2_n}}{n}}} \sim t_{n-1}$

if H_0 were true, this is the distribution of the test statistic. The null distribution.

(observed statistic) $\quad t^* \quad \leftarrow$ statistic the data gave you. if it's far in the tails of the null distribution, that's evidence against the null hypothesis.

(p-value) $\quad \text{p-val} = P(|T_n| \geq t^* \mid H_0 \text{ true})$



Stuff we sweep under the rug in AP Stat or 101

- derivation of sampling distributions
- why is it $n-1$ everywhere and not n?
- whence the t distribution?
- why are $\hat{\mu}_n$ and $\widehat{\sigma_n^2}$ independent?

We will tackle the first 3 points. We defer the fourth to STA 332. You need Calc III and linear algebra.

The tool kit

$$X \sim N(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$\implies cX \sim N(c\mu, c^2\sigma^2)$$

$$Z \sim N(0,1) \implies Z^2 \sim \chi^2_1 = \text{Gamma}\left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

$$X \sim \text{Gamma}(\alpha, \beta) \implies E(X) = \alpha/\beta$$

$$\text{var}(X) = \alpha/\beta^2$$

$$m(t) = E[e^{tX}] = \left(\frac{\beta}{\beta - t}\right)^\alpha \qquad t < \beta$$

$$cX \sim \text{Gamma}(\alpha, \beta/c)$$

$$X_1, \dots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta) \implies \sum_{i=1}^{n} X_i \sim \text{Gamma}(n\alpha, \beta)$$

$$\hat{\mu}_n \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right) \implies \frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \sim N(0,1)$$

Recall the Student's t derivation

independence $\left\{ \begin{array}{l} V \sim \text{Gamma}\left(\frac{df}{2}, \frac{df}{2}\right) \\ Z \sim N(0,1) \end{array} \right.$

$$T = Z/\sqrt{V}$$

$$\Downarrow$$

$$T \mid V = v \sim N(0, 1/v)$$

$$\implies$$

$$V \sim \text{Gamma}\left(\frac{df}{2}, \frac{df}{2}\right)$$
$$T \mid V = v \sim N(0, 1/v)$$

joint dist of (V, T) written hierarchically

$$T \sim t_{df}$$

Marginal distribution of T

Let's show: $\widehat{\sigma_n^2} \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2} \frac{1}{\sigma_0^2} \right)$

BTW: $$E\left(\widehat{\sigma_n^2} \right) = \frac{\cancel{\frac{n-1}{2}}}{\cancel{\frac{n-1}{2}} \frac{1}{\sigma_0^2}} = \sigma_0^2$$ unbiased!

$$\text{var}\left(\widehat{\sigma_n^2} \right) = \frac{\frac{n-1}{2}}{\left(\frac{n-1}{2} \frac{1}{\sigma_0^2} \right)^2} = \frac{4\sigma_0^2}{n-1} \to 0$$ consistent!

Itchy calculation:

$$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \mu_0 + \hat{\mu}_n - \hat{\mu}_n)^2$$

$$= \sum_{i=1}^n \left[(X_i - \hat{\mu}_n) + (\hat{\mu}_n - \mu_0) \right]^2$$

$$= \sum_{i=1}^n \left[(X_i - \hat{\mu}_n)^2 + 2(\hat{\mu}_n - \mu_0)(X_i - \hat{\mu}_n) + (\hat{\mu}_n - \mu_0)^2 \right]$$

$$= \sum_{i=1}^n (X_i - \hat{\mu}_n)^2 + 2(\hat{\mu}_n - \mu_0) \sum_{i=1}^n (X_i - \hat{\mu}_n) + \sum_{i=1}^n (\hat{\mu}_n - \mu_0)^2$$

$$= \sum_{i=1}^n (X_i - \hat{\mu}_n)^2 + 2(\hat{\mu}_n - \mu_0) \left[\sum_{i=1}^n X_i - \sum_{i=1}^n \hat{\mu}_n \right] + \sum_{i=1}^n (\hat{\mu}_n - \mu_0)^2$$

$$= \sum_{i=1}^n (X_i - \hat{\mu}_n) + 2(\hat{\mu}_n - \mu_0) \left[\sum_{i=1}^n X_i - n\hat{\mu}_n \right] + \sum_{i=1}^n (\hat{\mu}_n - \mu_0)^2$$

$$= \sum_{i=1}^n (X_i - \hat{\mu}_n) + 2(\hat{\mu}_n - \mu_0) \left[\sum_{i=1}^n X_i - n \frac{1}{n} \sum_{i=1}^n X_i \right] + \sum_{i=1}^n (\hat{\mu}_n - \mu_0)^2$$

$$= \sum_{i=1}^n (X_i - \hat{\mu}_n) + 2(\hat{\mu}_n - \mu_0) \left[\cancelto{0}{\sum_{i=1}^n X_i - \sum_{i=1}^n X_i} \right] + \sum_{i=1}^n (\hat{\mu}_n - \mu_0)^2$$

$$\sum_{i=1}^n (X_i - \mu_0)^2 = (n-1)\widehat{\sigma_n^2} + n(\hat{\mu}_n - \mu_0)^2$$

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu_0)^2 = \frac{(n-1)}{\sigma_0^2} \widehat{\sigma_n^2} + n \frac{(\hat{\mu}_n - \mu_0)^2}{\sigma_0^2}$$

$$\sum_{i=1}^n \frac{(X_i - \mu_0)^2}{\sigma_0^2} = \frac{(n-1)}{\sigma_0^2} \widehat{\sigma_n^2} + \frac{(\hat{\mu}_n - \mu_0)^2}{\sigma_0^2 / n}$$

$$\sum_{i=1}^n \left(\frac{x_i - \mu_0}{\sigma_0} \right)^2 = \frac{n-1}{\sigma_0^2} \widehat{\sigma_n^2} + \left(\frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma_0^2 / n}} \right)^2$$

Recall if B, C are independent random variables, then

$$M_{B+C}(t) = E\left[e^{t(B+C)}\right] = E\left[e^{tB+tC}\right] = E\left[e^{tB}e^{tC}\right] = E\left[e^{tB}\right]E\left[e^{tC}\right] = M_B(t)M_C(t)$$

independence!

$$\underbrace{\sum_{i=1}^{n} \underbrace{\underbrace{\left(\frac{X_i - \mu_0}{\sigma_0}\right)}_{N(0,1)}^2}_{\text{Gamma}(1/2,\,1/2)}}_{\text{Gamma}\left(\frac{n}{2},\,\frac{1}{2}\right)} = \underbrace{\frac{(n-1)}{\sigma_0^2}\widehat{\sigma_n^2}}_{???} + \underbrace{\underbrace{\left(\frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}}\right)}_{N(0,1)}^2}_{\text{Gamma}\left(\frac{1}{2},\,\frac{1}{2}\right)}$$

$$A \quad = \quad B \quad + \quad C$$

$$M_A(t) = M_{B+C}(t) = M_B(t)\,M_C(t) \implies M_B(t) = \frac{M_A(t)}{M_C(t)}$$

$$= \frac{\left(\frac{1/2}{1/2 - t}\right)^{n/2}}{\left(\frac{1/2}{1/2 - t}\right)^{1/2}}$$

$$= \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{n-1}{2}}$$

So $\quad \dfrac{(n-1)}{\sigma_0^2}\widehat{\sigma_n^2} \sim \text{Gamma}\left(\dfrac{n-1}{2}, \dfrac{1}{2}\right)$

$$\implies \widehat{\sigma_n^2} = \frac{\sigma_0^2}{n-1}\,\frac{n-1}{\sigma_0^2}\widehat{\sigma_n^2} \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}\,\frac{1}{\sigma_0^2}\right)$$

What happens if we use n instead of $n-1$?

$$\widehat{\sigma_n^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2} \frac{1}{\sigma_0^2}\right)$$

$$(n-1)\widehat{\sigma_n^2} = \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2} \frac{1}{\sigma_0^2}\right)$$

$$\frac{n-1}{n}\widehat{\sigma_n^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n}{2} \frac{1}{\sigma_0^2}\right)$$

So

$$E\left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2\right] = \frac{\frac{n-1}{2}}{\frac{n}{2}\frac{1}{\sigma_0^2}} = \underbrace{\frac{n-1}{n}\sigma_0^2 < \sigma_0^2}$$

- biased downward!
- for large n, doesn't matter too much

If you standerdize w/ the <u>true</u> variance . . .

$$\frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \sim N(0,1)$$

only one source of randomness

if you plug in the estimated variance, which is more realistic . . .

$$\frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\widehat{\sigma_n^2}}{n}}} \sim t_{n-1}.$$

two sources of randomness make tails heavier

$$\frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\widehat{\sigma_n^2}}{n}}} = \frac{\frac{\hat{\mu}_n - \mu_0}{1/\sqrt{n}}}{\sqrt{\widehat{\sigma_n^2}}} = \frac{1/\sigma_0}{1/\sigma_0} \; \frac{\frac{\hat{\mu}_n - \mu_0}{1/\sqrt{n}}}{\sqrt{\widehat{\sigma_n^2}}}$$

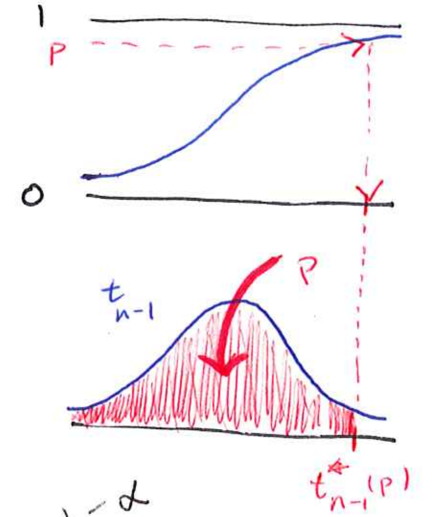
$$= \frac{\frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma_0^2/n}}}{\sqrt{\widehat{\sigma_n^2}/\sigma_0^2}} \quad \begin{matrix} N(0,1) \\ \updownarrow \text{ independent!} \\ \mathrm{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}\right) \end{matrix}$$

$$\sim t_{n-1}$$

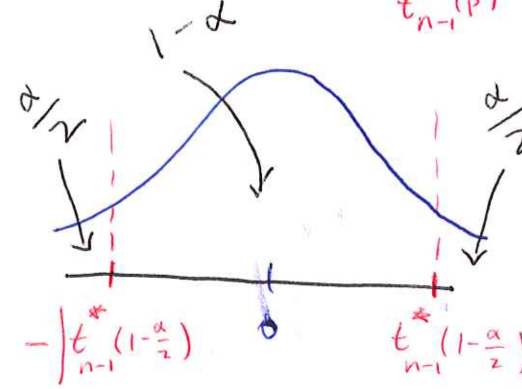
Building the exact CI for the mean of iid normal data

- For any sample size n, we have

$$\frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\widehat{\sigma_0^2}}{n}}} \sim t_{n-1} .$$



- Let $t^*_{n-1}(p)$ be quantile of t_{n-1} distribution
- We know for $\alpha = 0.01, 0.05, 0.1$, etc ...



$$P\left(-t^*_{n-1}(1-\tfrac{\alpha}{2}) < \frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\widehat{\sigma_n^2}}{n}}} < t^*_{n-1}(1-\tfrac{\alpha}{2}) \right)$$

exact!

$$= 1 - \alpha .$$

- $$P\left(-t^*_{n-1}(1-\tfrac{\alpha}{2}) < \frac{\hat{\mu}_n - \mu_0}{\sqrt{\frac{\widehat{\sigma_n^2}}{n}}} < t^*_{n-1}(1-\tfrac{\alpha}{2}) \right) = 1-\alpha$$

$$P\left(-t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} < \hat{\mu}_n - \mu_0 < t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} \right) = 1-\alpha$$

$$P\left(-\hat{\mu}_n - t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} < -\mu_0 < -\hat{\mu}_n + t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} \right) = 1-\alpha$$

$$P\left(\hat{\mu}_n + t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} > \mu_0 > \hat{\mu}_n - t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}} \right) = 1-\alpha$$

- So, if you take

$$L_n = \hat{\mu}_n - t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}}$$

$$U_n = \hat{\mu}_n + t^*_{n-1}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{\sigma_n^2}}{n}}$$

then $P(L_n < \mu_0 < U_n) = 1-\alpha$ For all n!

exact, not approximate