Statistics and Data Science for Engineers
E178 / ME276DS

Statistical inference: Confidence intervals and Hypothesis tests

Point estimation

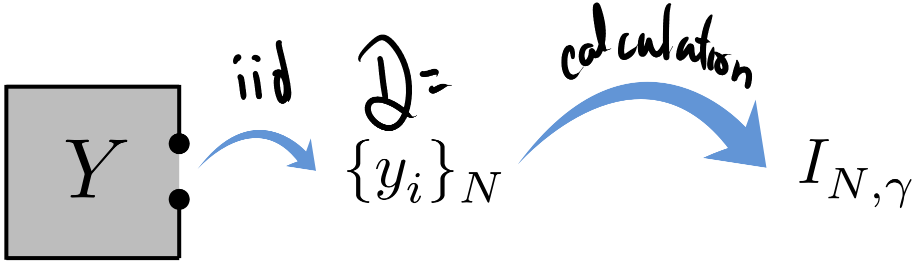
- 3 simple example: Sample mean
  Unbiased sample variance
  Biased sample variance
- MLE.
- Complicated example.

Gaussian Mixture (Hidden variable) $\longrightarrow$ MLE $\xrightarrow{\text{Optimization problem}}$ Numerial algorithm: Expectation Maximization $\xrightarrow{\text{Assumptions. } \varepsilon \to 0}$ K-means

$\gamma = 0.9.$

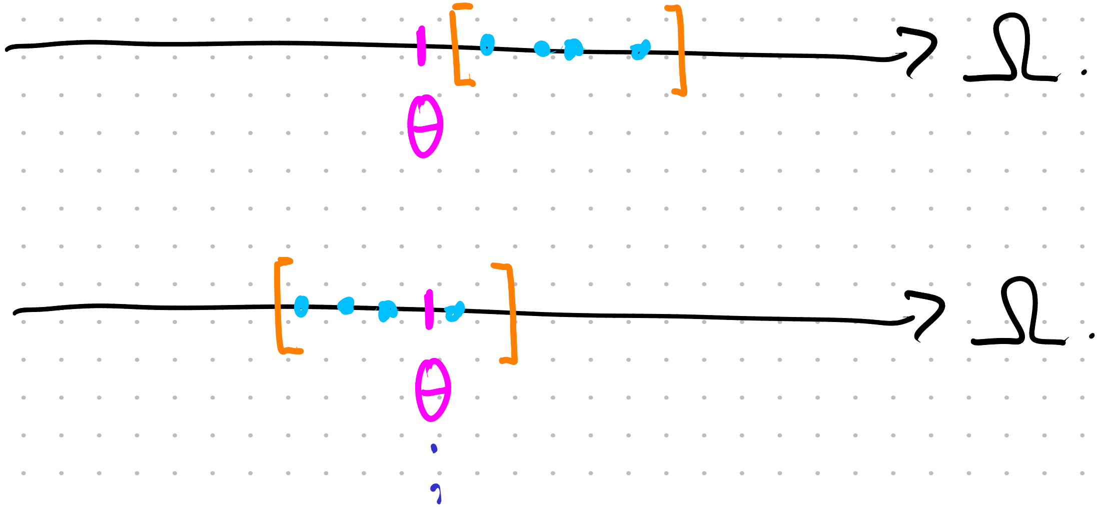
Confidence intervals

Statement: "I am 90% confident that $\theta \in I_{N,\gamma}$"



Meaning: "Repeating this procedure many times, the proportion of those in which $\theta \in I_{N,\gamma}$ tends to γ."

γ ... property of the calculation



$\theta \notin I_{N,\gamma}$: 0

$\theta \in I_{N,\gamma}$: 1

$\vdots$

$\text{mean}(0, 1, \dots) \to \gamma$

How do we calculate such an interval $I_{N,\gamma}$?

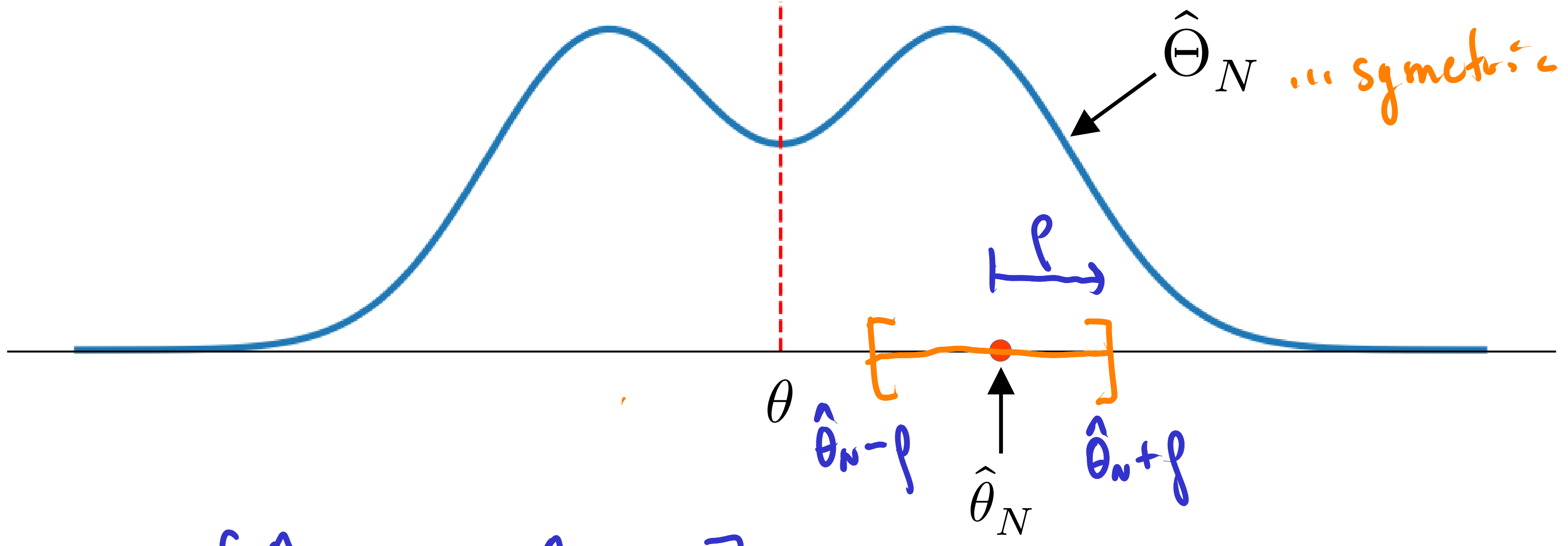
Assumption:

1. Have g_N an unbiased estimator of θ

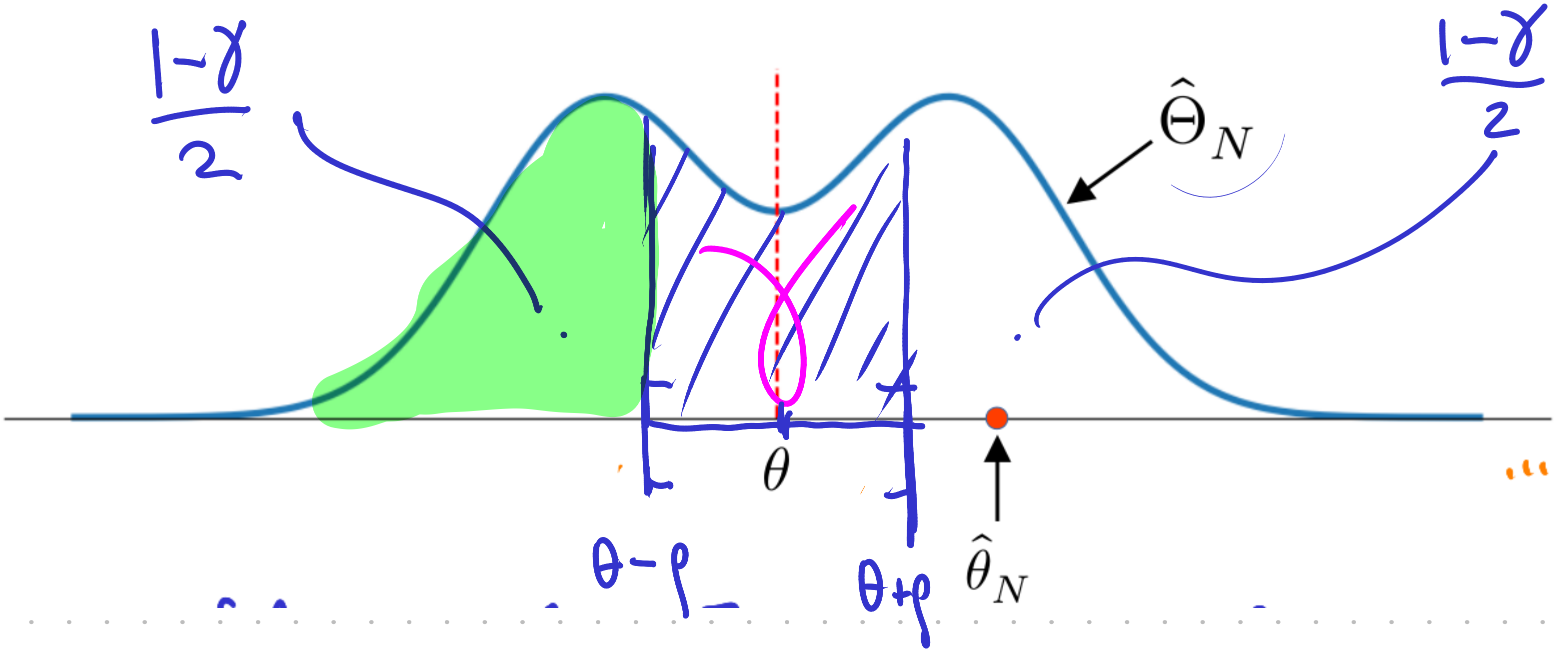
$$\{y_i\}_N \xrightarrow{g_N} \hat{\theta}_N.$$

$$\{Y_i\}_N \xrightarrow{g_N} \hat{\Theta}_N.$$

2. $\hat{\Theta}_N$ is symmetric.



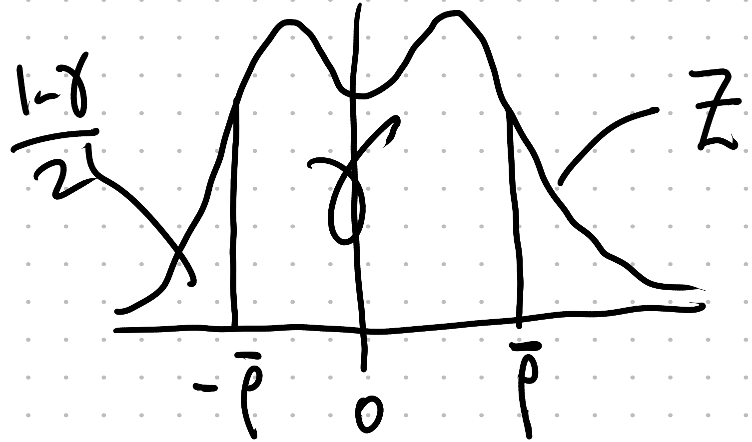
$$I_{N,\gamma} = [\hat{\theta}_N - \rho, \hat{\theta}_N + \rho].$$



"Repeating this procedure many times, the proportion of those in which ~~$\theta \in I_{N,\gamma}$~~ $\hat{\Theta}_N \in I'$ tends to γ." $= P_{\hat{\Theta}_N}(I')$

$$\to \Phi_{\hat{\Theta}_N}(\theta - \rho) = \frac{1-\gamma}{2} \quad \text{I}$$

Problem: I don't know θ



Solution: Normalization..

$$Z = \frac{\hat{\Theta}_N - \theta}{\sigma_{\hat{\Theta}_N}}$$

$$\bar{\rho} = \frac{\rho}{\sigma_{\hat{\Theta}_N}}$$

I become: $\Phi_Z(-\bar{\rho}) = \frac{1-\gamma}{2}$.

$$-\bar{\rho} = \Phi_Z^{-1}\left(\frac{1-\gamma}{2}\right)$$

$\gamma \in [0,1]$.

$$\bar{\rho} = \left|\Phi_Z^{-1}\left(\frac{1-\gamma}{2}\right)\right|$$

$$\therefore \boxed{\rho = \sigma_{\hat{\Theta}_N}\left|\Phi_Z^{-1}\left(\frac{1-\gamma}{2}\right)\right|}$$

Summary: Confidence interval for θ

Problem: Given

1. $\mathcal{D} = \{y_i\}_N \overset{\text{iid}}{\sim} Y$ ...
2. An unbiased estimator of θ : $\hat{\Theta}_N = g_N(Y_1, \dots, Y_N)$
3. $\hat{\Theta}_N$ is symmetric

find an interval $I_{N,\gamma}$ that contains θ with confidence γ.

Solution: $I_{N,\gamma} = \hat{\theta}_N \pm \rho$ with

$$\hat{\theta}_N = g_N(y_1, \dots, y_N)$$

$$\rho = \sigma_{\hat{\Theta}_N} \left| \Phi_Z^{-1}\left(\frac{1-\gamma}{2}\right) \right|$$

Z = normalized $\hat{\Theta}_N$

Confidence interval for the mean of a normal distribution (σ_Y known)

Problem: Given

1. $\mathcal{D} = \{y_i\}_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ = Y
2. $\bar{Y}_N$ is an unbiased estimator of μ_Y (sample mean).
3. $\bar{Y}_N = \mathcal{N}(\mu_Y, \sigma_Y^2/N)$ is symmetric ... $\hat{\Theta}_N$

$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$

find an interval $I_{N,\gamma}$ that contains μ_Y with confidence γ.

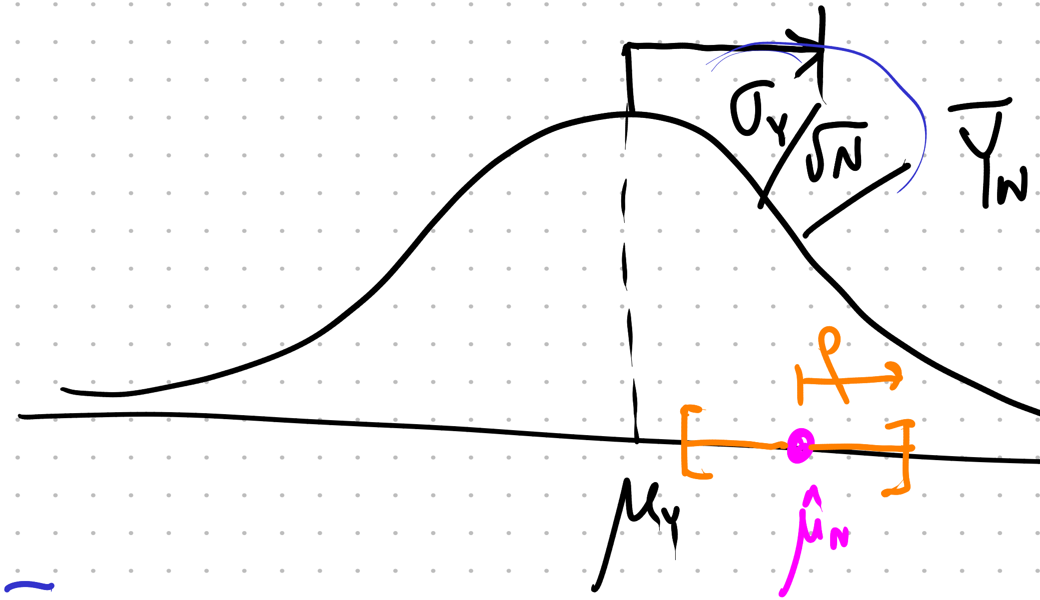
Solution: $I_{N,\gamma} = \hat{\mu}_N \pm \rho$ with

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\rho = \frac{\sigma_Y}{\sqrt{N}}\left|\Phi_{\mathcal{N}}^{-1}\left(\frac{1-\gamma}{2}\right)\right|$$

$$Z = \mathcal{N}(0,1)$$

$Z = \mathcal{N}(0,1)$



... unbiased

$$E[\bar{Y}_N] = \mu_Y$$

$$Z = \frac{\bar{Y}_N - \mu_Y}{\sigma_Y/\sqrt{N}} = \mathcal{N}(0,1).$$

$$\bar{\rho}$$

$$\rho = \frac{\sigma_Y}{\sqrt{N}} \boxed{\left| \Phi_N^{-1}\left(\frac{1-\gamma}{2}\right) \right|}$$

Example Process Y is Gaussian with $\sigma_Y = 3.$

Find a 90% confidence interval for μ_Y using the following data.

```
D = array([10.53834891, 14.78970656, 11.89977148,  4.77130213,  9.14813745,
           15.69801573, 14.95192423, 12.10085195,  2.65706828, 13.77066178,
           12.19813675, 10.76307477,  5.14012851,  7.03212572, 13.79827886])
```

Solution

```
gamma = 0.9
N = D.shape[0]
muhatN = D.mean()
sigmaY = 3
```

$= 15$ $\hat{\mu}_N$

import scipy.stats as stats.

$ppf = \Phi^{-1}$

```
rhobar = abs(stats.norm().ppf((1-gamma)/2))
rhobar
```

$N(0,1)$

1.6448536269514729

```
rho = (sigmaY/np.sqrt(N))*rhobar
rho
```

1.2740981408263843

```
ci = [muhatN-rho, muhatN+rho]
ci
```

[9.343070733411258, 11.891267015064026]

- μ_Y is in $[9.34, 11.89]$ with 90% confidence
- $\mu_Y = 10.61 \pm 1.27$ with 90% confidence.

Using tables:

```
(1-gamma)/2
```

```
0.049999999999999
```

0.05

```
invcdf = -1.645
rho = sigmaY/np.sqrt(N) * abs(invcdf)
rho
```

```
1.2742115209022402
```

Lookup table

x	$\Phi_{\mathcal{N}}^{-1}(x)$
0.001	-3.090
0.0025	-2.807
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645
0.1	-1.282
0.15	-1.036
0.2	-0.842
0.25	-0.674
0.3	-0.524
0.4	-0.253

Confidence interval for the mean of a normal distribution (σ_Y unknown)

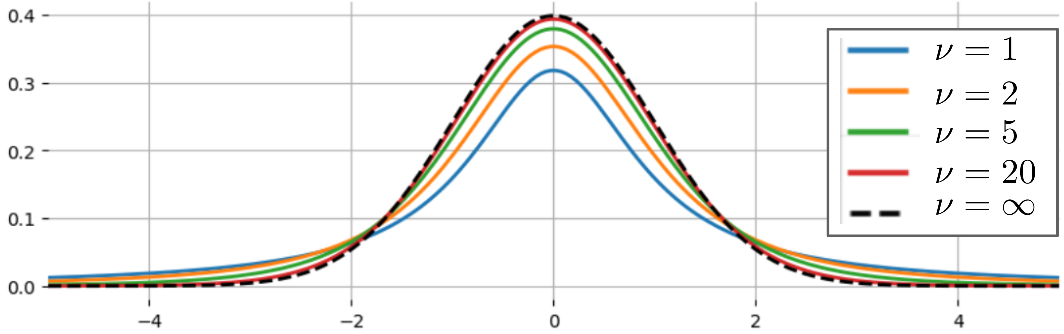
- Solution: Estimate σ_Y from the data. ... using the unbiased sample variance.

$$\hat{\sigma}_N = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{\mu}_N)^2}$$

- Instead of $\cancel{Z = \frac{\bar{Y}_N - \mu_Y}{\sigma_Y/\sqrt{N}}}$ use $t = \frac{\bar{Y}_N - \mu_Y}{S_N/\sqrt{N}}$

Degrees of freedom.

The t distribution $(\nu = N-1)$



N is large : $t \to N$

N is small : $t \neq N$

S_N^2 ... unbiased sample variance

$\hat{\sigma}_N^2 \sim S_N^2$

Example Process Y is Gaussian ~~with $\sigma_Y = 3$~~.
Find a 90% confidence interval for μ_Y using the same data as before.

Solution

```
gamma = 0.9
N = D.shape[0]
muhatN = D.mean()
sigmahatN = D.std(ddof=1)
```

$$\frac{1}{N-\text{ddof}} \sum_{i=1}^{N} (y_i - \hat{\mu}_N)^2$$

```
tdist = stats.t(df=N-1)
rhobar = abs(tdist.ppf((1-gamma)/2))
rhobar
```

$\nu = N-1$

$\Phi_t^{-1}\left(\frac{1-\gamma}{2}\right)$

1.7613101357748566

```
rho = sigmahatN/np.sqrt(N) * rhobar
rho
```

1.847271343874377

$$\boxed{\rho = \frac{\hat{\sigma}_N}{\sqrt{N}} \left| \Phi^{-1}_{t(N-1)}\left(\frac{1-\gamma}{2}\right) \right|}$$

```
ci = [muhatN-rho, muhatN+rho]
ci
```

[8.769897530363265, 12.464440218112019]

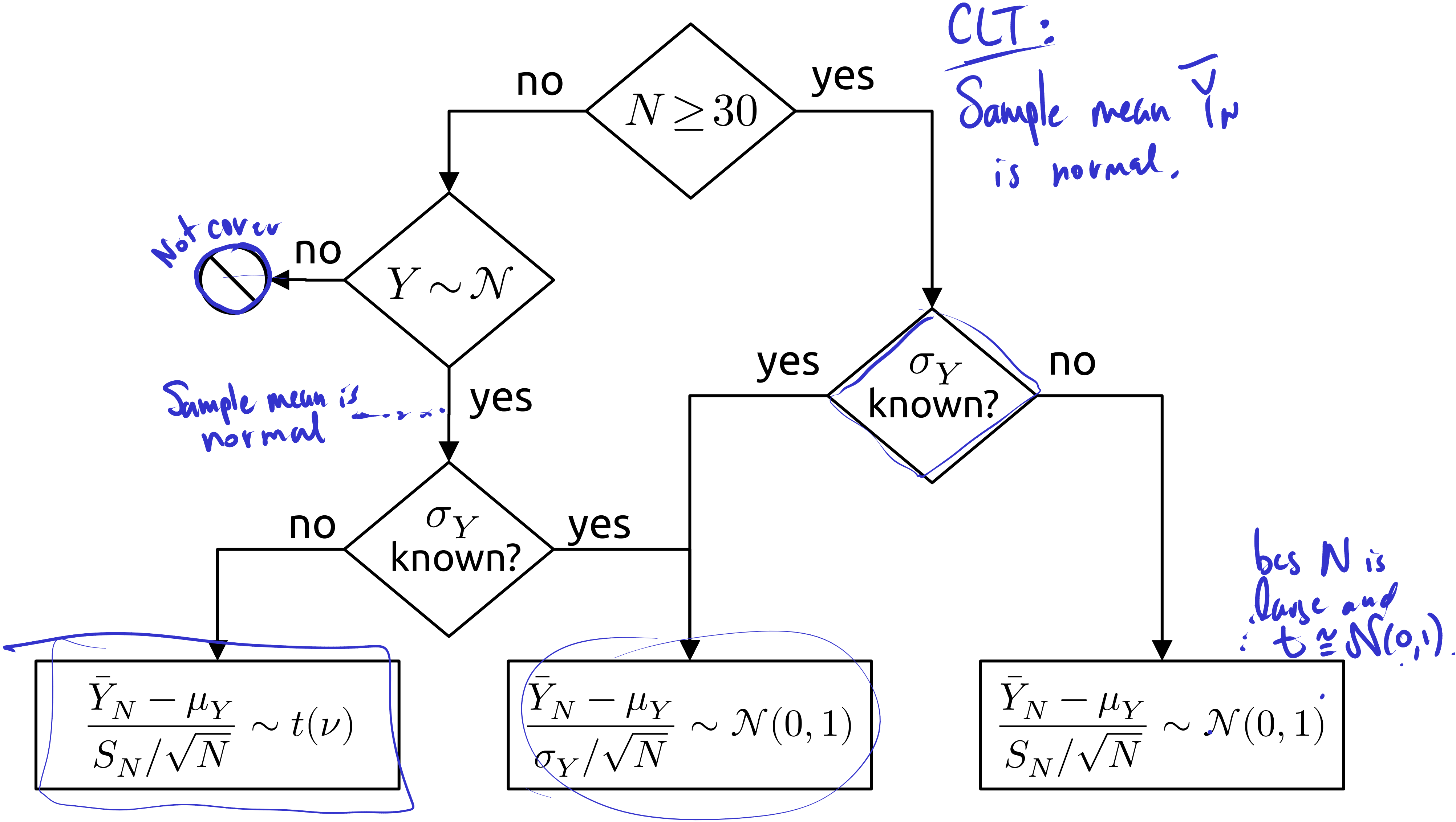
Using tables:

$N = 15$ $\quad \gamma = 0.9 \rightarrow \frac{1-\gamma}{2} = 0{,}05$ $\quad \Phi^{-1}_{t(\nu)}(x)$

```
invcdf = -1.761
rho = sigmahatN/np.sqrt(N) * abs(invcdf)
rho
```

1.8469460718408115

x	$\nu=1$		$\nu=13$	$\nu=14$
0.001	-318.3		-3.852	-3.787
0.0025	-127.3		-3.372	-3.326
0.005	-63.65		-3.012	-2.977
0.01	-31.82		-2.650	-2.624
0.025	-12.70		-2.160	-2.145
0.05	-6.314	•••	-1.771	-1.761
0.1	-3.078		-1.350	-1.345
0.15	-1.963		-1.079	-1.076
0.2	-1.376		-0.870	-0.868
0.25	-1.000		-0.694	-0.692
0.3	-0.727		-0.538	-0.537
0.4	-0.325		-0.259	-0.258



$N \geq 30$
no
yes
CLT:
Sample mean $\bar{Y}_N$ is normal.
$Y \sim \mathcal{N}$
Not cover
Sample mean is normal
σ_Y known?
σ_Y known?
$\frac{\bar{Y}_N - \mu_Y}{S_N/\sqrt{N}} \sim t(\nu)$
$\frac{\bar{Y}_N - \mu_Y}{\sigma_Y/\sqrt{N}} \sim \mathcal{N}(0,1)$
$\frac{\bar{Y}_N - \mu_Y}{S_N/\sqrt{N}} \sim \mathcal{N}(0,1)$
bcs N is large and $t \cong \mathcal{N}(0,1)$

Confidence interval for the mean of a Bernoulli r.v.

Problem: Given

1. $\mathcal{D} = \{y_i\}_N \overset{\text{iid}}{\sim} \mathcal{B}(p)$ $\quad \mathcal{D} = \{1, 0, 0, 1\}$ $\hat{p} = \frac{2}{4}$
2. $\bar{Y}_N$ is the sample mean (unbiased estimator of p)
3. $N\,\bar{Y}_N = \mathcal{B}in(N, p)$

$\hat{\mu}_N = \frac{1}{N} \sum y_i$

$\bar{Y}_N = \frac{1}{N} \sum Y_i$

$N\bar{Y}_N = \sum Y_i = \#$ ones.

find an interval $I_{N,\gamma}$ that contains p with confidence γ.

Assume $N > 30$.

Solution: $I_{N,\gamma} = \hat{p} \pm \rho$ with

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{N^+}{N}$$

$$\rho = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}} \left| \Phi_{\mathcal{N}}^{-1}\left(\frac{1-\gamma}{2}\right) \right|$$

$$Z = \mathcal{N}(0,1)$$

Side note:

$\mathcal{B}in \to \mathcal{N}$
as $N \to \infty$
by CLT.

$$\hat{\sigma}_{\hat{\Theta}_N} = \frac{\hat{\sigma}_N}{\sqrt{N}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}}.$$

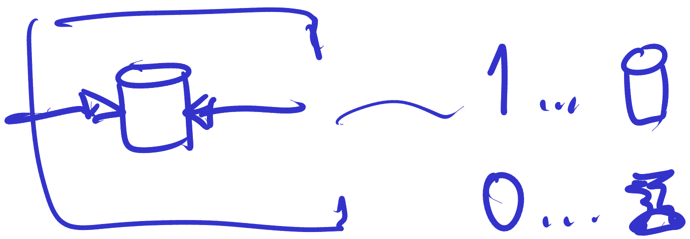
$$\hat{\sigma}_N^2 = \hat{p}(1-\hat{p}).$$

$$Y \sim B(p)$$

$$\mathrm{Var}[Y] = p(1-p).$$

Example (from Navidi ch. 5.2)

A soft-drink manufacturer purchases aluminum cans from an outside vendor. A random sample of 70 cans is selected from a large shipment, and each is tested for strength by applying an increasing load to the side of the can until it punctures. Of the 70 cans, 52 meet the specification for puncture resistance. Find a 95% confidence interval for the proportion of cans in the shipment that meet the specification.



$N = 70$, $\gamma = 0.95$.

$N^+ = 52$.

$$I_{N,\gamma} = \hat{p} \pm \rho \qquad \rho = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}} \left| \Phi^{-1}\left(\frac{1-\gamma}{2}\right) \right|$$

$$\hat{p} = \frac{N^+}{N} = \frac{52}{70} = 0.74$$

```python
gamma = 0.95
N = 70
Np = 52
phat = Np/N
rho = np.sqrt(phat*(1-phat)/N) * abs(stats.norm.ppf((1-gamma)/2))
rho
```

```
0.10238561784264202
```

```python
I = [phat-rho,phat+rho]
I
```

```
[0.6404715250145009, 0.8452427606997849]
```

$I = 0.74 \pm 0.1$ with 95% confidence.

Confidence intervals.

1. Assumption about distribution of Y.
   - Gaussian
   - Bernoulli : $\sigma_Y = \sqrt{p(1-p)}$
     $\hat{\sigma}_N = \sqrt{\hat{p}(1-\hat{p})}$
   - ⋮

2. θ parameter
   - $\mu_Y \; (E[Y])$.
   - point est. $\sigma_Y^2 \; (Var[Y])$.

3. C.I. on <u>mean</u>. ....

   Estimator : $\bar{Y}_{1N}$.
   Estimate : $\hat{\mu}_N$

   $\longrightarrow I = \hat{\mu}_N \pm \rho$.

4. $$\rho = \frac{\sigma_Y}{\sqrt{N}} \left| \Phi^{-1}_{N}\left(\frac{1-\gamma}{2}\right) \right|.$$

Variant #1 → Known ... σ_Y

→ unknown ... $\hat{\sigma}_N$. ⟶ $\mathcal{N}$ becomes $t(\nu)$.

But! When $N > 30$

Then we can use CLT.

to assert $\bar{Y}_N$ is Gaussian

regardless of the distribution of Y

Hypothesis tests

$\theta \ldots \mu$

Given: Null hypothesis: $H_0 : \theta = x$

Alternate hypothesis: $H_1 : \theta \neq x$ or $\theta < x$ or $\theta > x$

Choose between two statements

- H_0 is rejected in favor of H_1.
- H_0 is *not* rejected in favor of H_1.

Types of Error

- $\boxed{\text{Type I:}}$ H_0 is rejected when it is true. ... put an innocent person in jail
- Type II: H_0 is not rejected when it is false... letting a criminal go free.

H_0 ... defendant.
H_1 ... prosecutor

1. Presume innocence.
   (Assume H_0 is true)

2. Collect evidence. : D.

3. Estimate likelihood of evidence assuming innocense.

4. If the evidence is unlikely <u>and</u> supports the prosecutor's case
   $\Rightarrow$ guilty. (Reject H_0 in favor of $H1$).

$$\boxed{Y \atop E[Y]=250} \sim \{248, 249, 250\}.$$

$H_0: \mu_Y = 250$

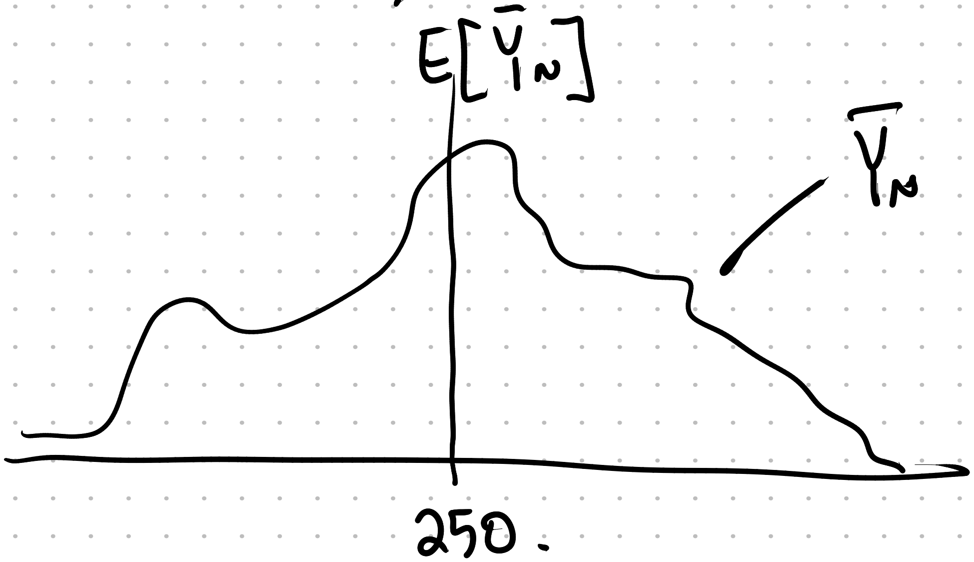
$H_1: \mu_Y < 250.$

0. Assume H_0 is true.

1. Sample mean estimator $\longrightarrow$ unbiased.

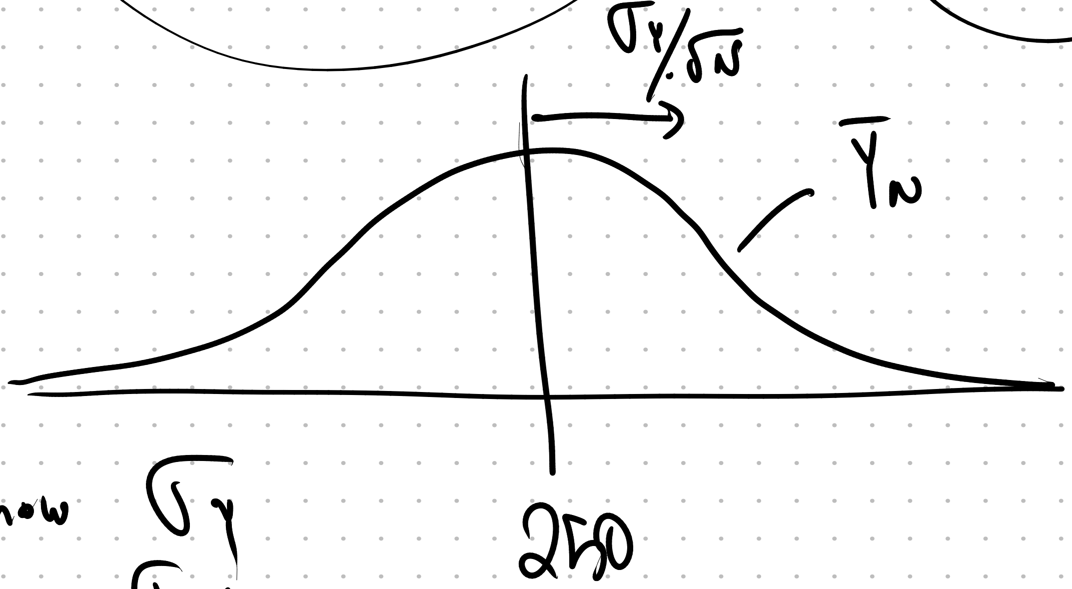
$$\hat{\mu}_N = \frac{1}{3}(248 + 249 + 250) = \underline{249}$$

$$\bar{Y}_N = \frac{1}{3} \sum_{i=1}^{3} Y_i$$



2. Assume $\bar{Y}_N$ is Gaussian.

Y is Gaussian or. $N > 30$.



3. Assume know σ_Y

$\Rightarrow \sigma_{\bar{Y}_N} = \sigma_Y/\sqrt{N}$

Example Given $\mathcal{D} = \{248, 249, 250\} \overset{\text{iid}}{\sim} Y$ test: $H_0 : \mu_Y = 250$

$H_1 : \mu_Y < 250$

$Y = \mathcal{N}(\mu_Y, 25)$ $\leftarrow \sigma_Y^2$

Assumption 2 ..

$\alpha = 0.05$

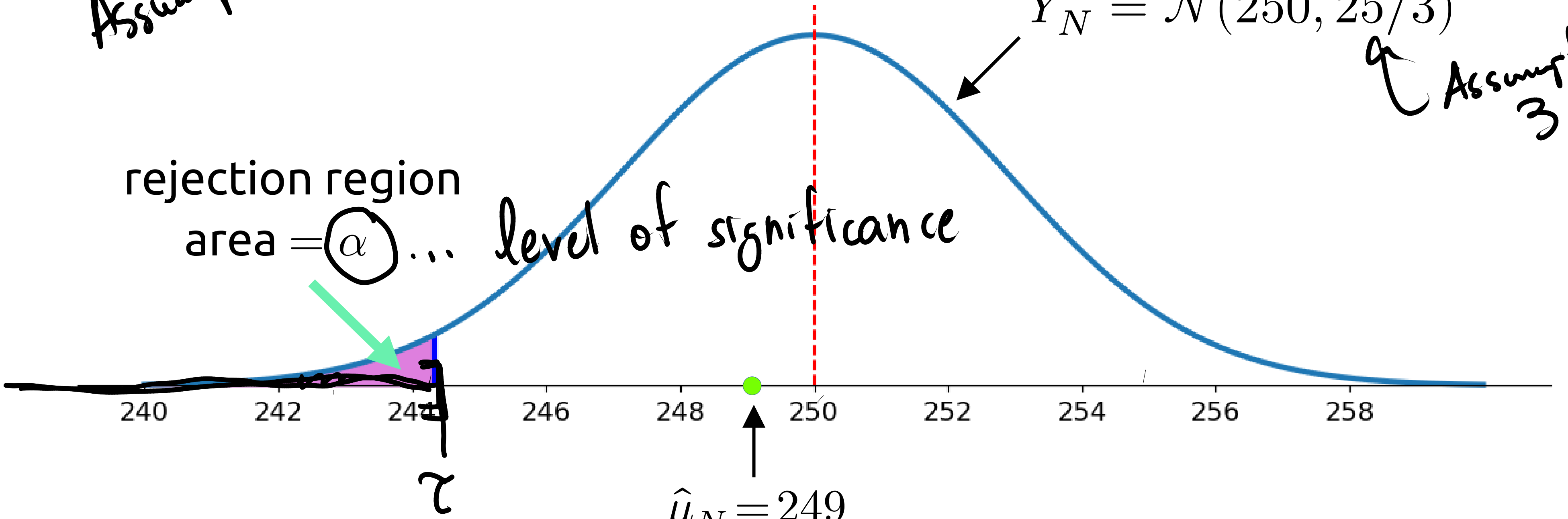
Assumption 0.

$\bar{Y}_N = \mathcal{N}(250, 25/3)$

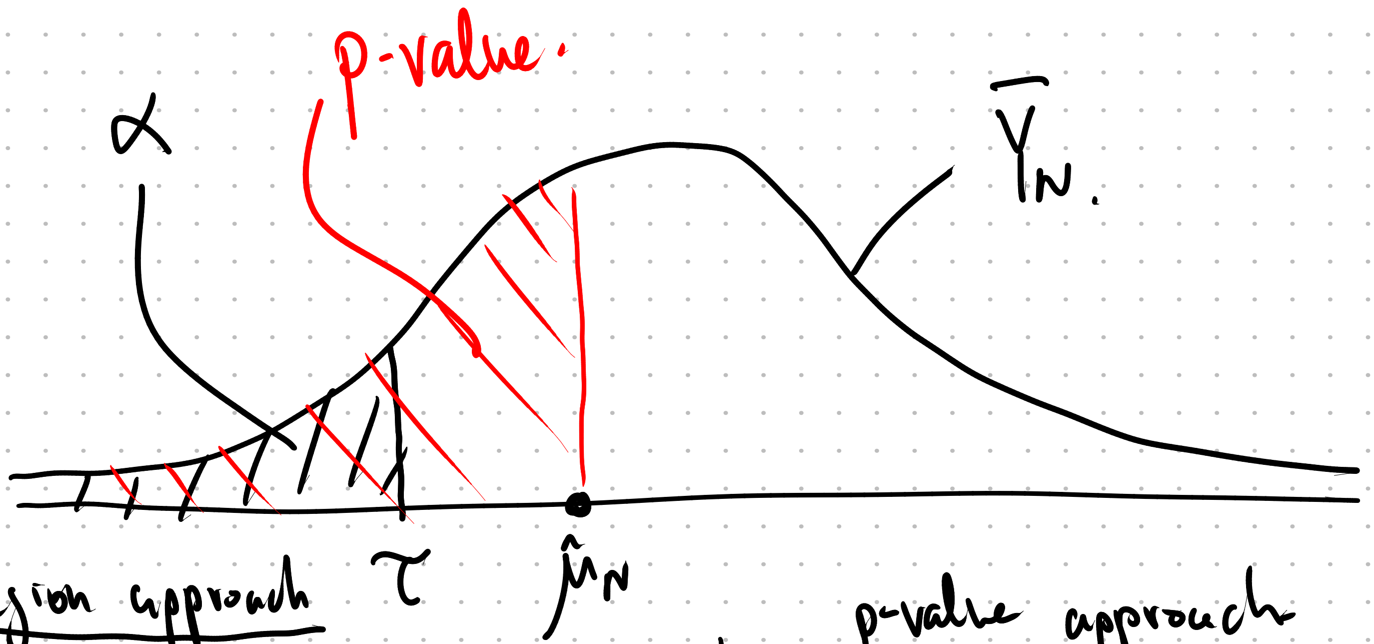
Assumption 3.

rejection region

area $= \alpha$ ... level of significance



$\hat{\mu}_N = 249$



<u>rejection region approach</u>

$$\Phi_{\bar{Y}_N}(\tau) = \alpha$$

$$\therefore \quad \tau = \Phi^{-1}_{\bar{Y}_N}(\alpha).$$

Reject $H_0 \iff \hat{\mu}_N < \tau$

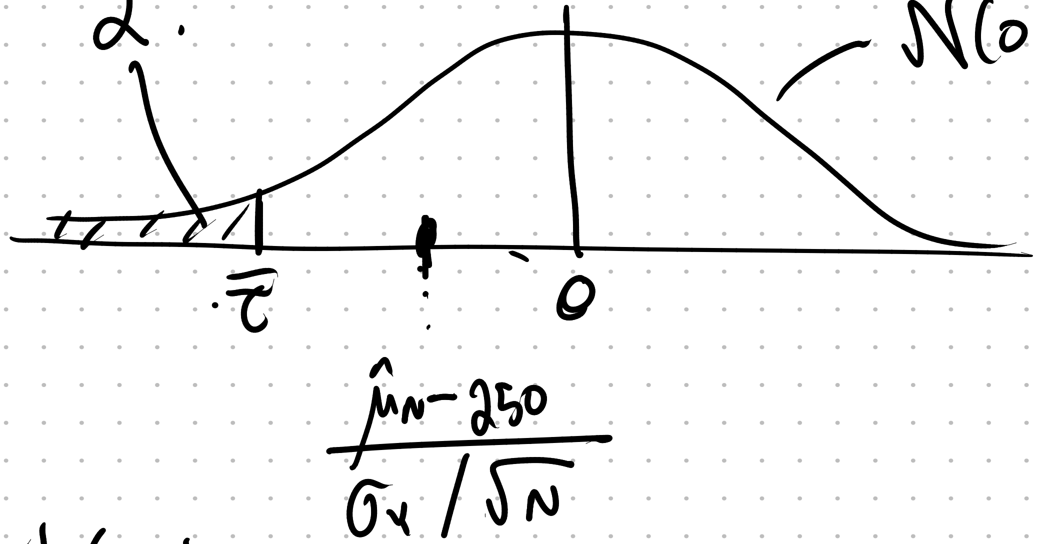
p-value approach

$$\text{p-value} = \Phi_{\bar{Y}_N}(\hat{\mu}_N).$$

Reject $H_0 \iff \text{pvalue} < \alpha$.

Normalize:

$$Z = \frac{\bar{Y}_N - 250}{\sigma_Y / \sqrt{N}} \sim \mathcal{N}(0,1).$$



$$\bar{\tau} = \Phi_{\mathcal{N}}^{-1}(\alpha).$$

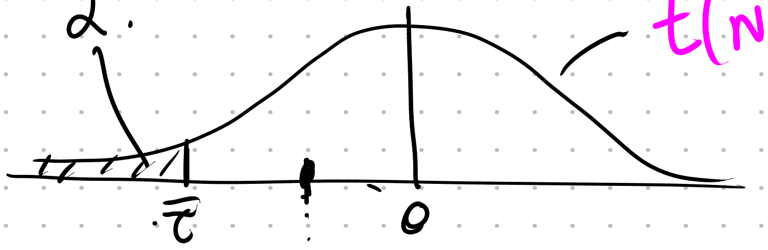
$$\text{Reject } H_0 \iff \frac{\hat{\mu}_N - 250}{\sigma_Y / \sqrt{N}} < \bar{\tau}$$

Case σ_Y is unknown (and $N < 30$) ... otherwise use Gaussian.

otherwise Gaussian

(and Y is Gaussian) ... required for using t distribution

$$t = \frac{\bar{Y}_N - 250}{\hat{\sigma}_N / \sqrt{N}} \sim t(\nu = N-1).$$



$$\frac{\hat{\mu}_N - 250}{\hat{\sigma}_N / \sqrt{N}}$$

$$\bar{\tau} = \Phi^{-1}_{t(N-1)}(\alpha).$$

$$\text{Reject } H_0 \iff \frac{\hat{\mu}_N - 250}{\hat{\sigma}_N / \sqrt{N}} < \bar{\tau}$$

General procedure

0\.

1. Choose a significance level α (e.g. $\alpha = 0.05$).
2. Use the unbiased estimator to compute $\hat{\theta}_N$.
3. Assume H_0 is true. Normalize the estimator and the estimate.
4. Use tables or software to find the rejection region. or p-value.
5. Determine whether the estimate falls within the rejection region.

or .... p value.

Example Given $\mathcal{D} = \{248, 249, 250\} \overset{\text{iid}}{\sim} Y$ test: $\boxed{\begin{array}{l} H_0 : \mu_Y = 250 \\ H_1 : \mu_Y < 250 \end{array}}$

$N = 3$

$Y = \mathcal{N}(\mu_Y, 25)$ $\sigma_Y = 5$

1. $\alpha = 0.05$

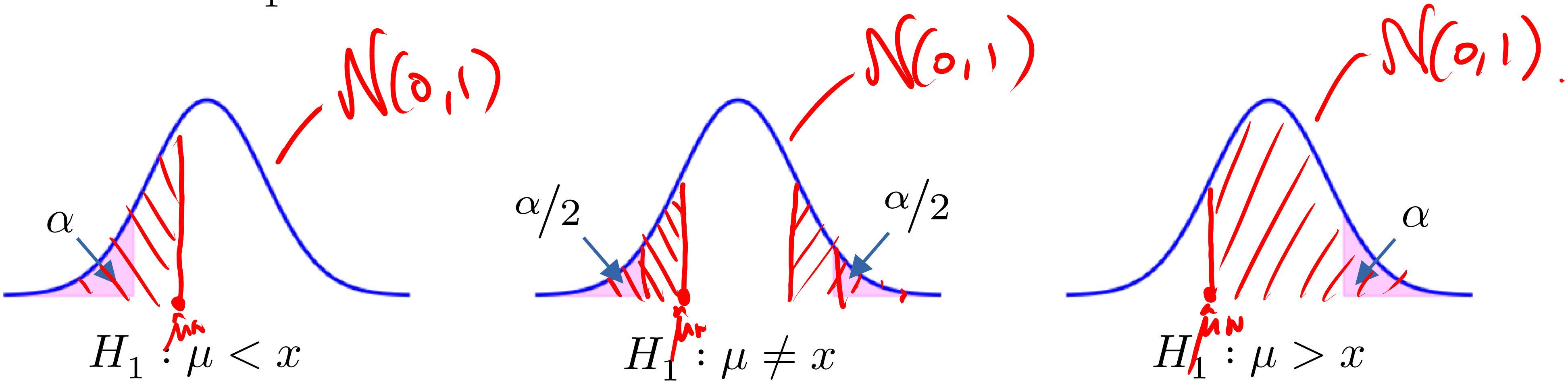
2. $\hat{\mu}_N = 249$ is a sample from $\bar{Y}_N$

3. $\frac{249-250}{5/\sqrt{3}} = -0.346$ is a sample from $Z = \frac{\bar{Y}_N - \mu_Y}{\sigma_Y/\sqrt{N}}$

4. $\tau = \Phi_Z^{-1}(0.05) = -1.645$ rejection region: $(-\infty, -1.645]$

5. $-0.346 \notin (-\infty, -1.645)$ $\Rightarrow$ H_0 is not rejected.

Possible H_1's



$$p = \Phi_N\left(\frac{\hat{\mu}_N - x}{\sigma_Y/\sqrt{N}}\right) \qquad p = 2\Phi_N\left(\frac{\hat{\mu}_N - x}{\sigma_Y/\sqrt{N}}\right) \qquad p = 1 - \Phi_N\left(\frac{\hat{\mu}_N - x}{\sigma_Y/\sqrt{N}}\right)$$

Reject H_0 .....H_1 $\iff$ $\boxed{\text{pvalue} < \alpha.}$

Variations

1. H.T. on the mean of a normal distribution, *unknown* σ_Y.
   - $N > 30 \Rightarrow$ use $\mathcal{N}(0,1)$
   - $N < 30 \Rightarrow$ use $t(\nu)$
2. H.T. on the mean of a Bernoulli distribution. $N > 30$.

Example (Navidi 8.10)

8.10 The hardness of a certain rubber (in degrees Shore) is claimed to be 65. Fourteen specimens are tested, resulting in an average hardness measure of 63.1 and a standard deviation of 1.4. Is there sufficient evidence to reject the claim, at the 5% level of significance? What assumption is necessary for your answer to be valid?

average

```
N = 14
muhat = 63.1
sigmahat = 1.4
alpha = 0.05

tau = abs(stats.t(df=N-1).ppf(alpha/2))
tau
```

```
2.160368656461013
```

```
t = (muhat-65)/(sigmahat/np.sqrt(N))
t
```

```
-5.07796359633606
```

```
pvalue = 2*stats.t(df=N-1).cdf(t)
pvalue
```

```
0.00021177635911378\53
```

$\mathcal{D} = \{ \ldots \}_{14}$

$\alpha = 0.05$.

$\hat{\mu}_N = 63.1$

$\hat{\sigma}_N = 1.4 \ \ldots$

$\sigma_Y = ?$

$H_0 : \mu_Y = 65$.

$H_1 : \mu_Y \neq 65$

$N = 14$

small.

σ_Y unknown.

N is small $\longrightarrow$ t distribution.

$\boxed{Y \text{ is Gaussian}}$

$$t = \frac{\hat{\mu}_N - 0}{\hat{\sigma}_N / \sqrt{N}}$$

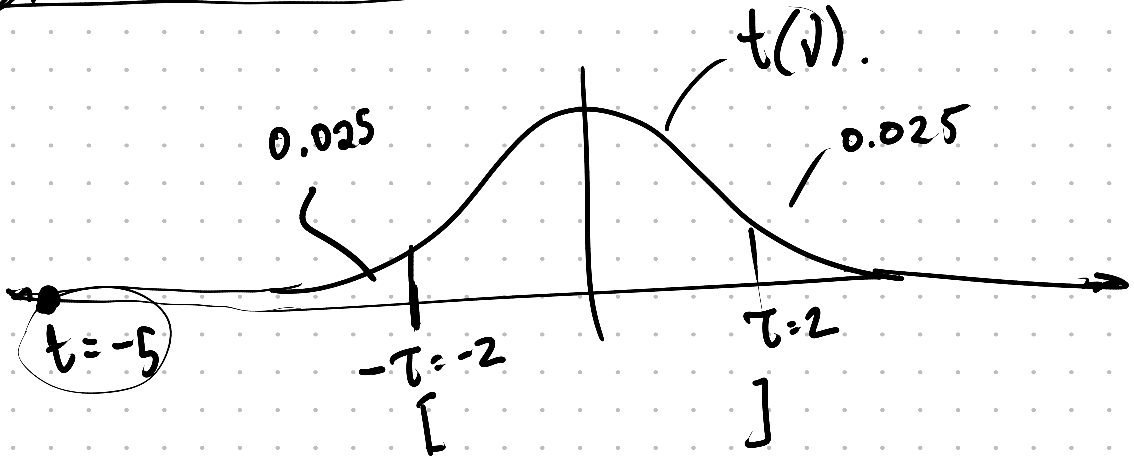
$$Z = \frac{\hat{\mu}_N - 0}{\sigma_Y / \sqrt{N}}.$$

1. Normalize

$$\boxed{t = \frac{\hat{\mu}_N - 65}{\hat{\sigma}_N / \sqrt{N}} = \underline{-5.}}$$

$$-\tau = \Phi^{-1}_{t(\nu=1)}(0.025)$$

Don't reject $\iff$ $|t| < \tau$



P-value: $2\Phi_t(t). \approx 0.0002.$