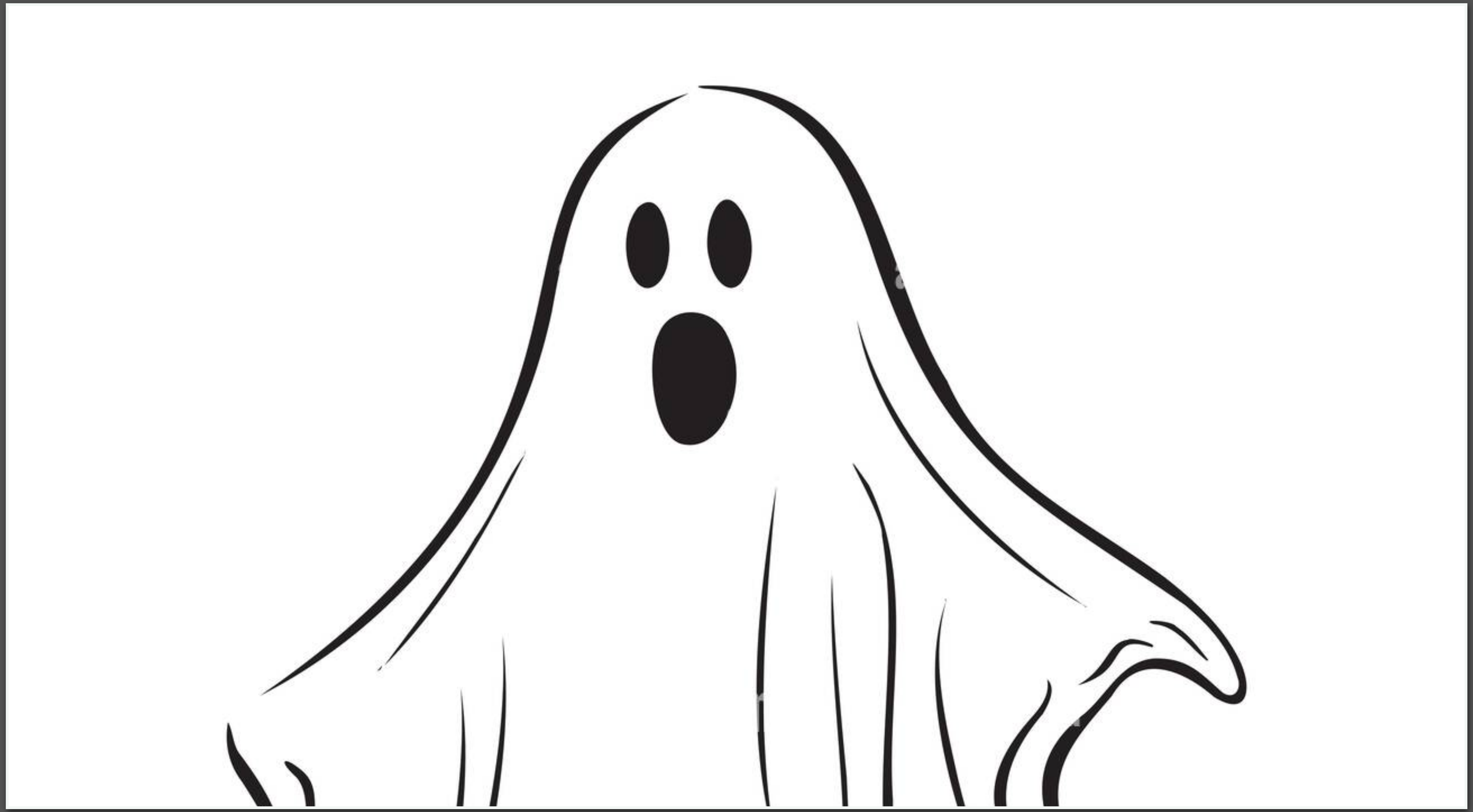


How normal is Normal Distribution?

And some other friendly distributions!!

Sandhya Tripathi
Staff Scientist, Department of Anesthesiology
Washington University School of Medicine in St Louis
October 25th, 2024

Formal name: Gaussian Distribution

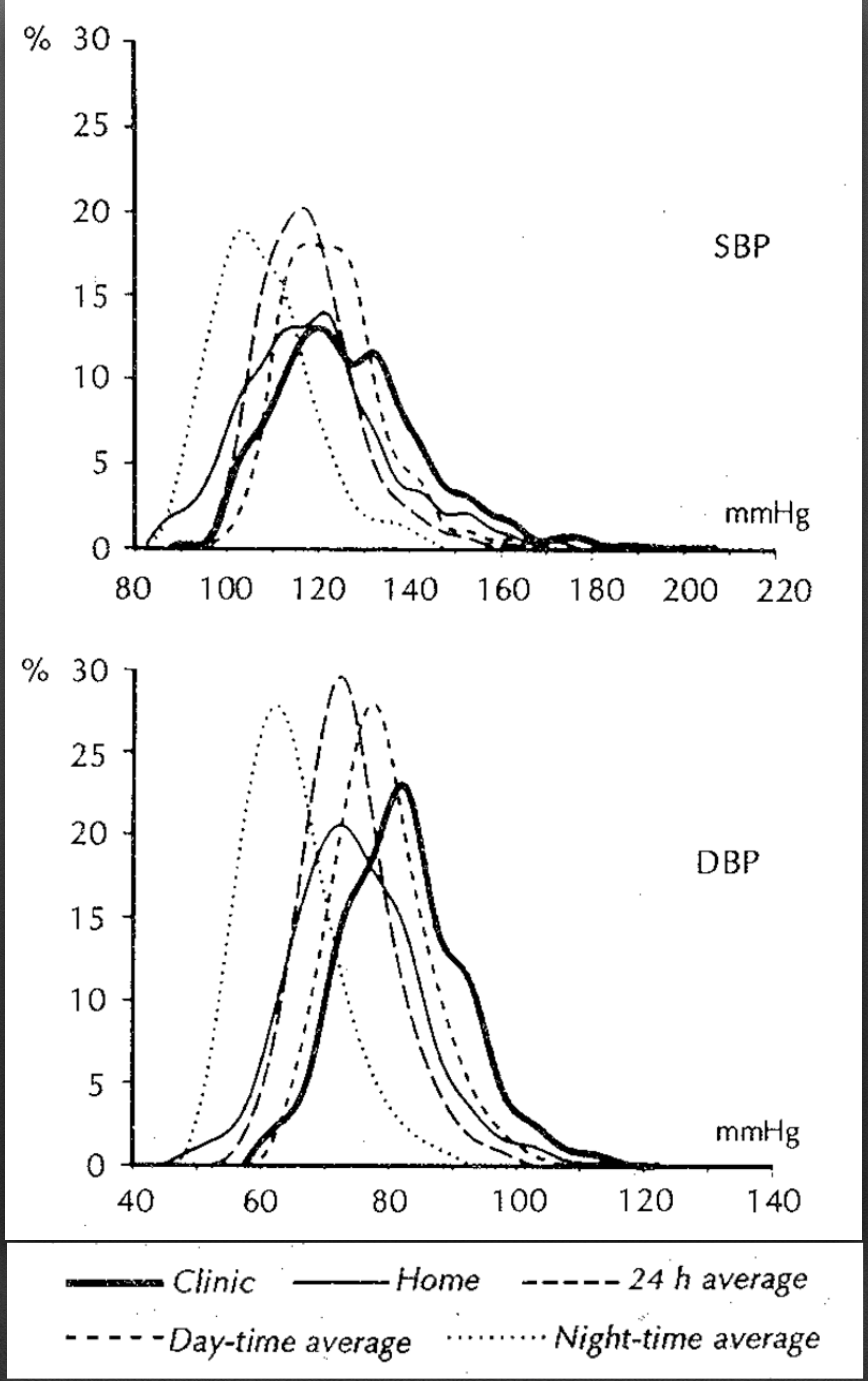
- Real valued continuous random variable: $X \sim N(\mu, \sigma^2)$
- Probability Density Function: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right)$$
- Cumulative Distribution Function: $F(x) = P[X \leq x]$
- Standard Normal Distribution: $Z \sim N(0,1)$ $\quad Z = \frac{X - \mu}{\sigma}$

$$\mathrm{P}[X \leq x] = \mathrm{P}\left[\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right] = \mathrm{P}\left[Z \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z)$$

This is what makes a lot of magic happen!

Blood pressure distribution in a general population

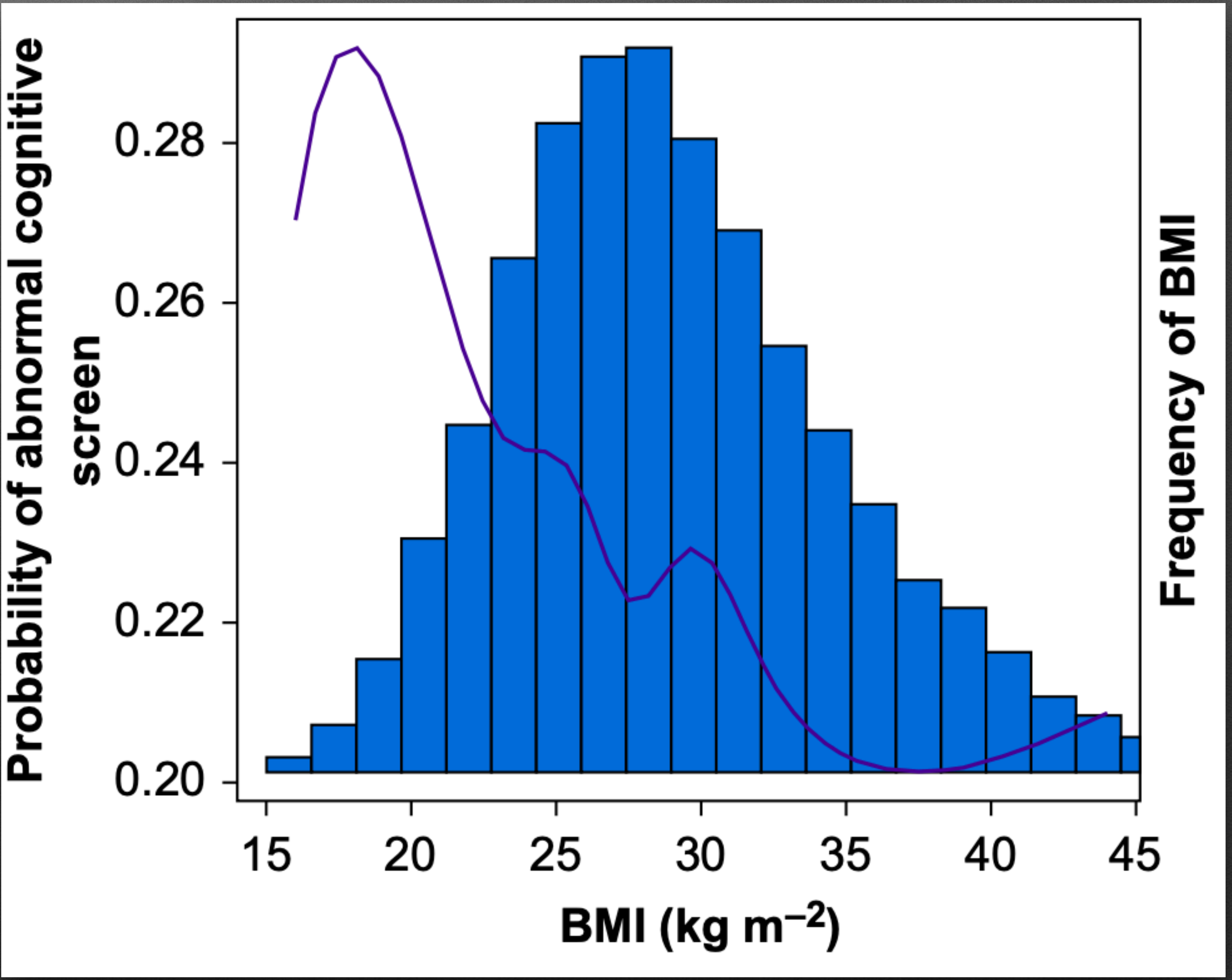
- Randomly selected 2400 patients representative of both sex and ages from the city of Manzo, Italy with population of 60000
- Normal blood pressure range in 1995: 140/90 mmHg
- Measured Systolic/Diastolic Blood Pressure in Clinic, Home, and Ambulatory (outpatient) settings
- <u>Findings: </u>120-130 and 75-81 mmHg which is close to the 'standard' upper limit



Source: Mancia, G., Sega, R., Bravi, C., De Vito, G., Valagussa, F., Cesana, G. and Zanchetti, A., 1995. Ambulatory blood pressure normality: results from the PAMELA study. *Journal of hypertension*, *13*(12), pp.1377-1390.

Body Mass Index of a large patient population

- Abnormal cognitive screening before surgery is known to be associated with postoperative complications
- 21666 patients >= 65 yr old
- Routine cognitive screening of aged surgical patients may be preferable to selective screening based on underlying comorbidities or other characteristics.



Source: Gregory, S.H., King, C.R., Abdallah, A.B., Kronzer, A. and Wildes, T.S., 2021. Abnormal preoperative cognitive screening in aged surgical patients: a retrospective cohort analysis. *British journal of anaesthesia*, *126*(1), pp.230-237.

TyG Index as a proxy for insulin resistance

- TyG: Tryglycerides and Fasting glycemia
- Low (TyG < 8.65) or high (TyG ≥ 8.65) risk for Type 2 Diabetes
- 256 patients (30% high risk)
- Goal was to develop a machine learning algorithm that predicts the future TyG values based on past
- Tested Normal distribution using Kolmogorov Smirnov Test

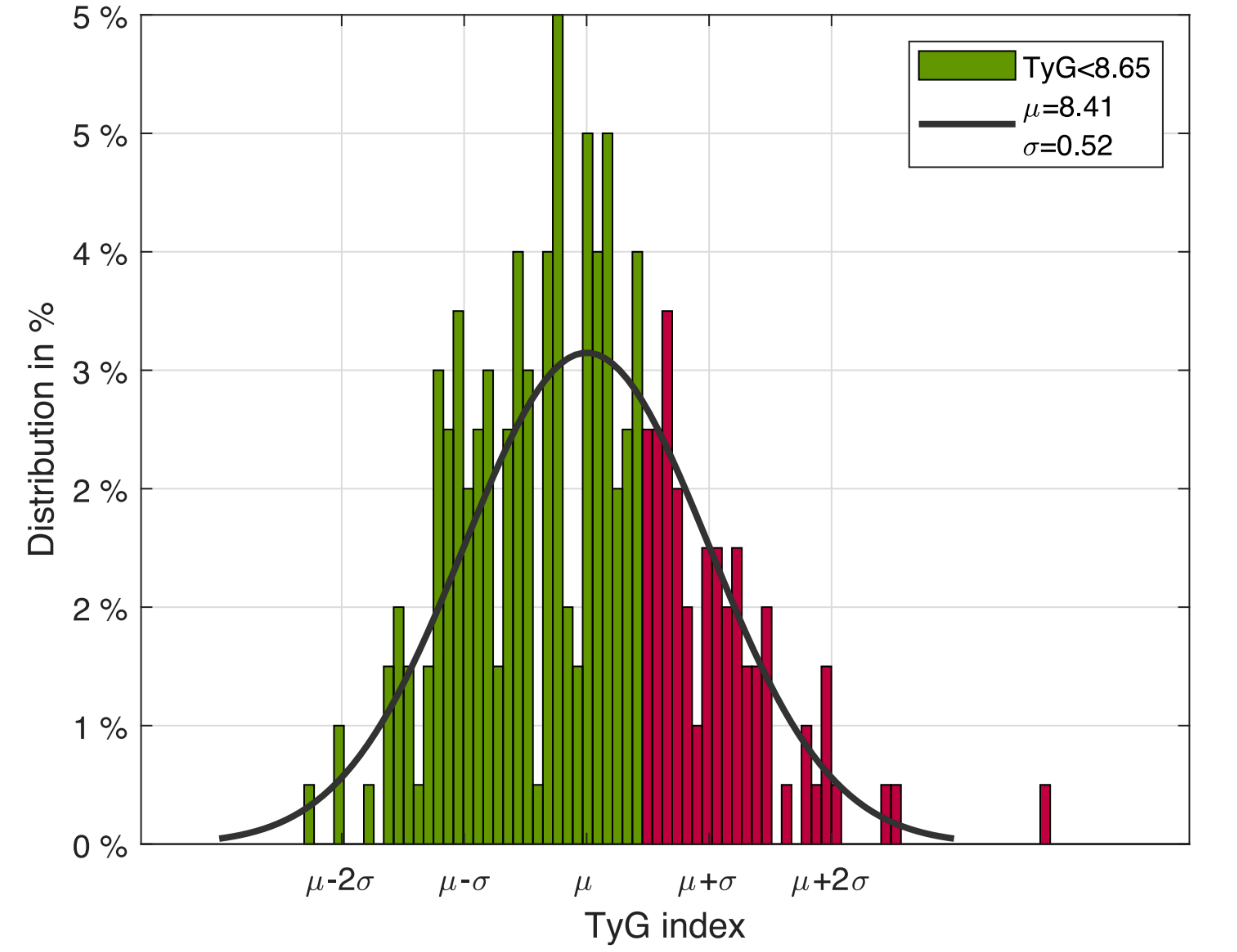


Fig. 5. TyG_{it_i} index distribution with mean $\mu = 8.41$ and standard deviation $\sigma = 0.52$. TyG index threshold ($\text{TyG}_{th} = 8.65$) separates the green side (179 patients) from the red side (77 patients) of the graph.

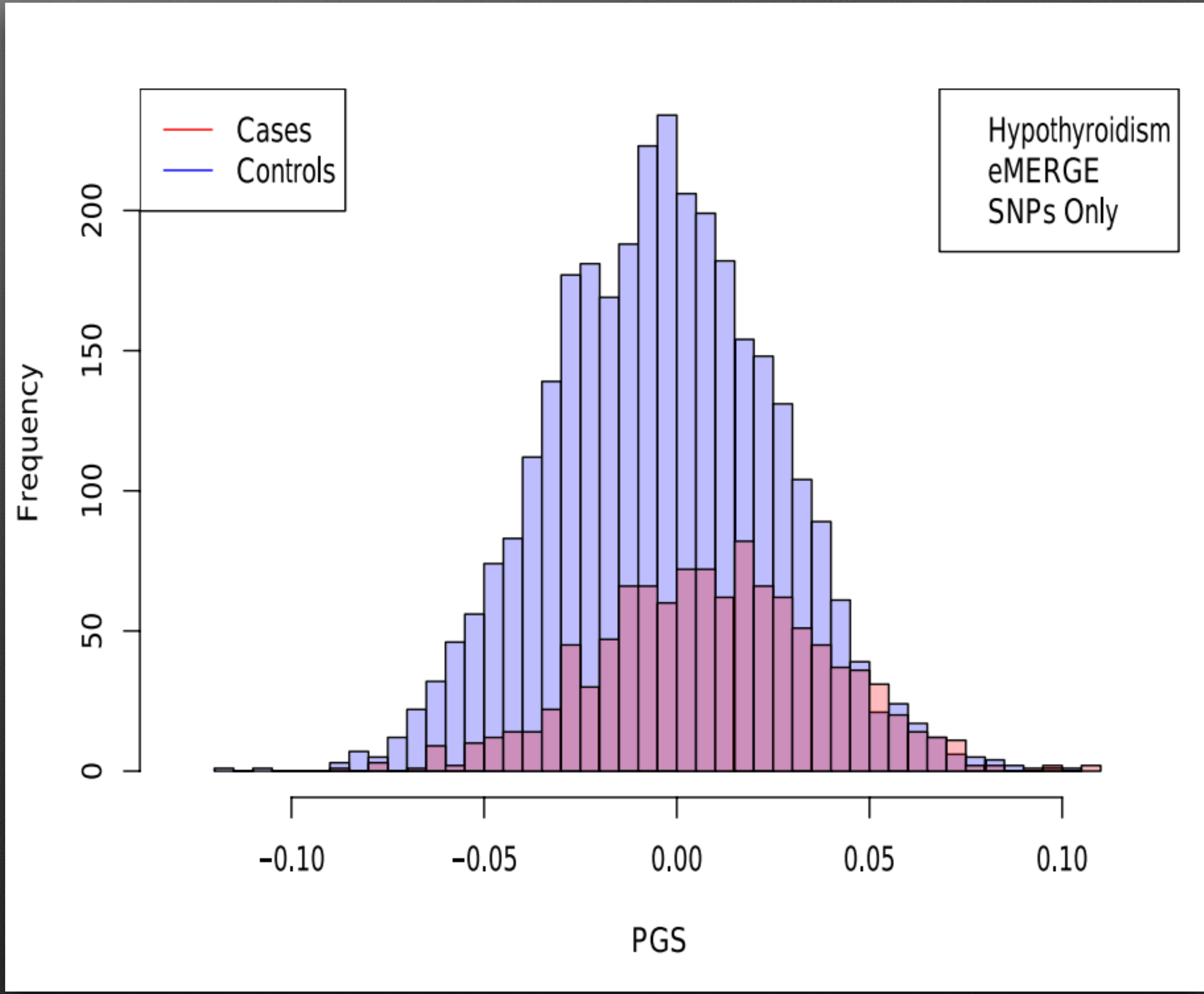
Source: Bernardini, M., Morettini, M., Romeo, L., Frontoni, E. and Burattini, L., 2020. Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach. *Artificial Intelligence in Medicine*, *105*, p.101847.

Derived Polygenic Score as disease predictors

- Polygenic Score: summarizes the estimated effect of many genetic variants on an individual's phenotype or possibilities of diseases

- A derived measure using single nucleotide polymorphisms (~genetic variations) which are further used as predictors of complex diseases

- UK Biobank with ~500,000 people



Source: Lello, L., Raben, T.G., Yong, S.Y., Tellier, L.C. and Hsu, S.D., 2019. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific reports*, *9*(1), p.15286.

Linear Regression with closed form solution

$$\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{X} + \epsilon$$

- Solution method: Ordinary least squares (OLS)
- Closed form solution for the coefficient is Best Linear Unbiased Estimate (BLUE)
- But how do we trust this estimate? Statistical Hypothesis Testing

Where does normality assumption fall and its implications?

- Need a test statistic that contains our parameter estimate information and is well characterized
- Assumption is $Y|X \sim N(E(Y|X), \sigma^2)$
- Equivalent to $\epsilon \sim N(0,\sigma^2)$
- These are the residuals, difference between the true and estimated y
- Testing for the normality of residuals is the deal!!

How do you check normality?

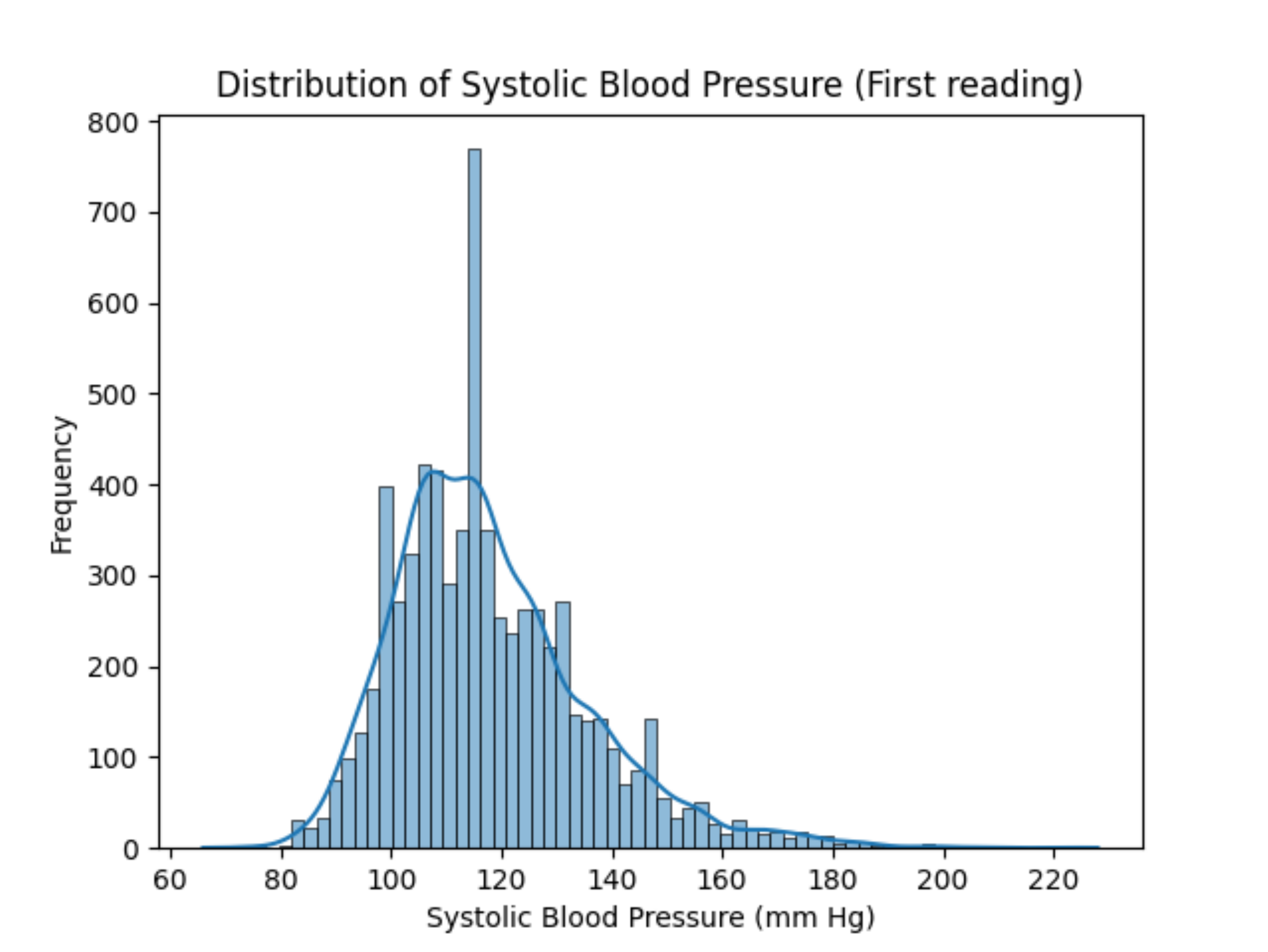
- Histograms of the residuals
- Sample quantiles vs Theoretical Distribution quantiles (Q-Q plots of residuals)
- Statistical Tests
    - Shapiro-Wilk test specifically for normality
    - Kolmogorov Smirnov one sample test

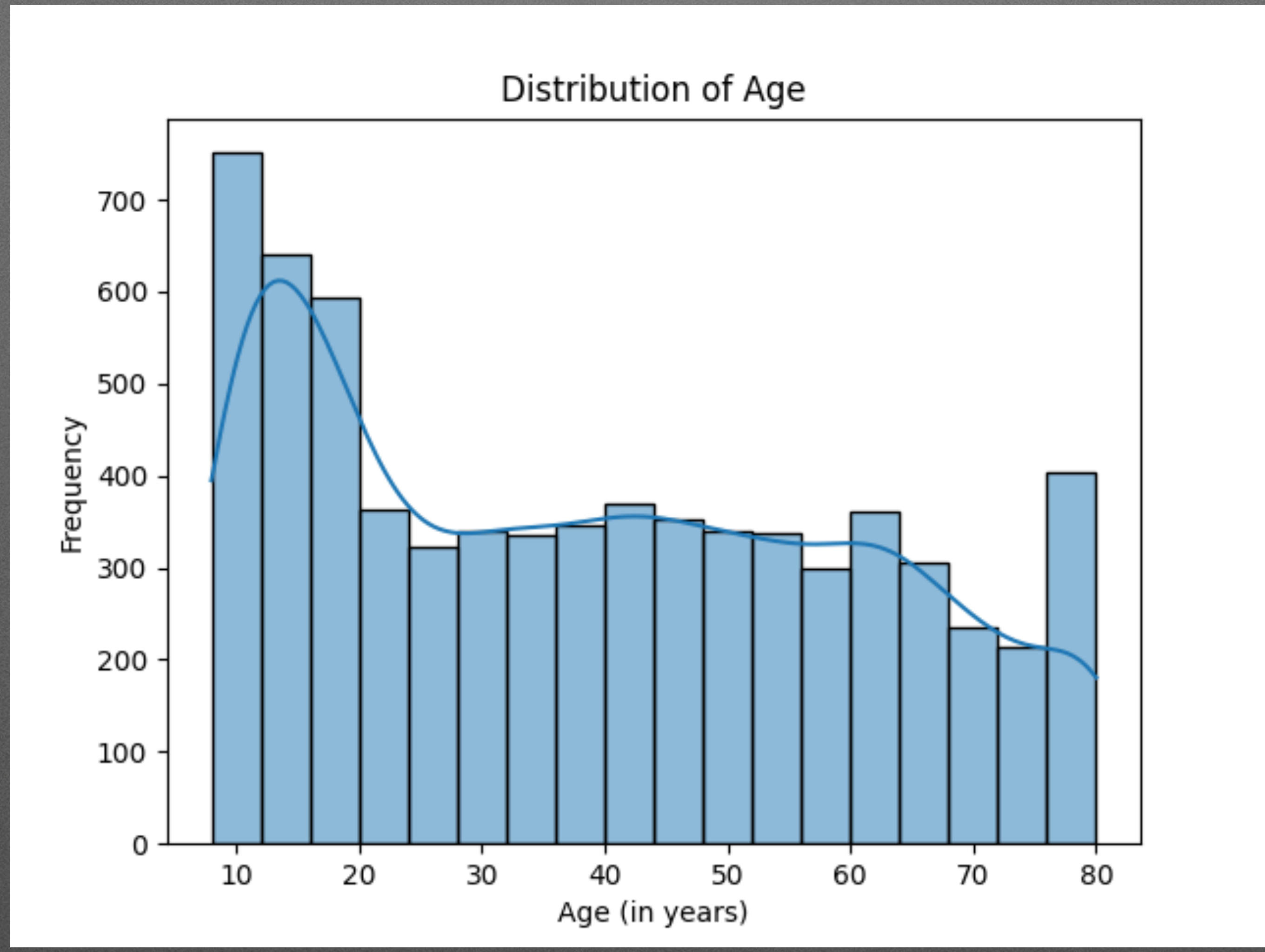
Demo on CDC-NHANES dataset

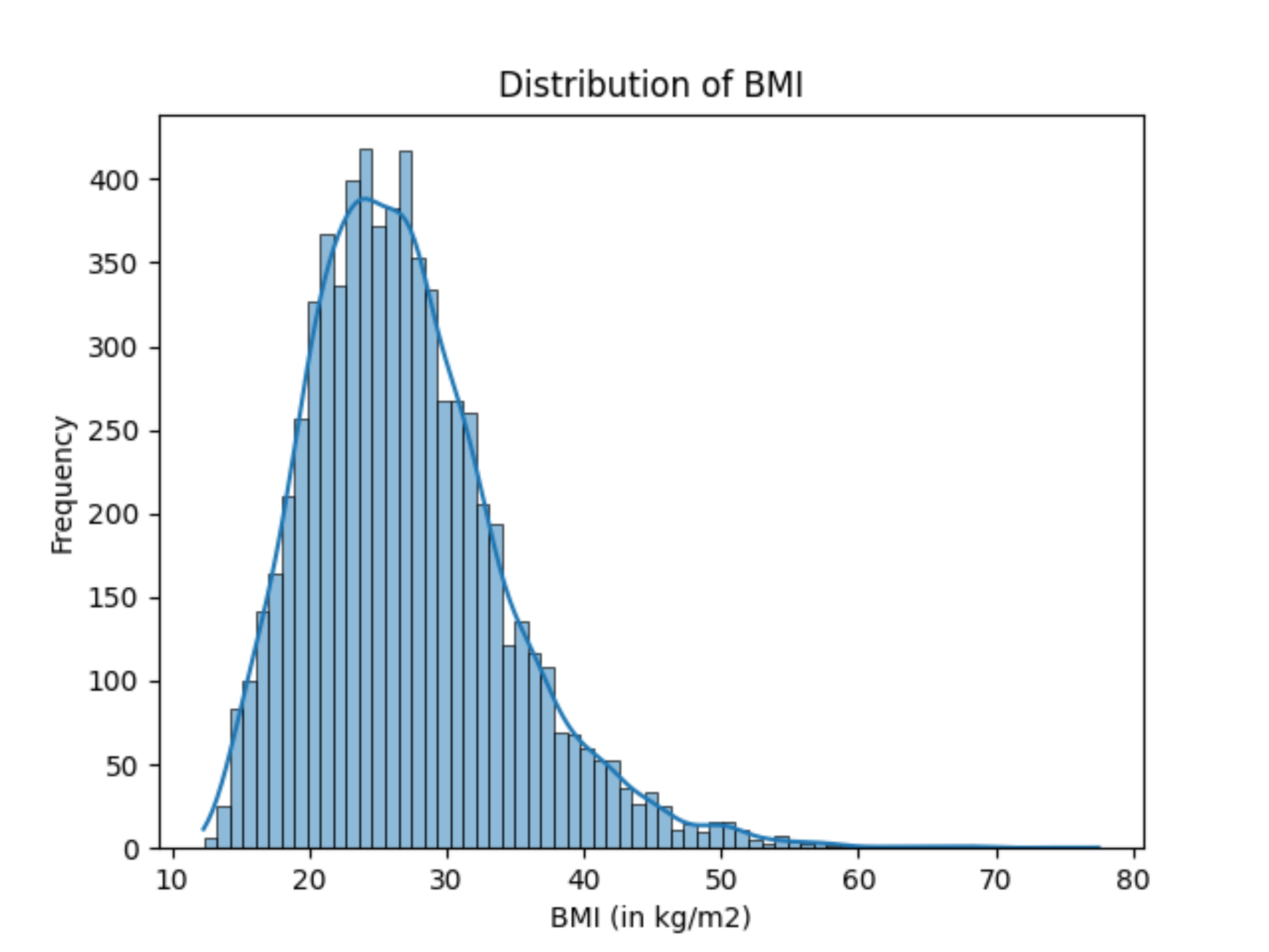
- Dataset Info: Health and nutritional status of adults and children in the US via interviews and examinations.
- Sample size: 6909 (2013-2014)
- Goal: Predict systolic blood pressure (SBP) based on age and BMI using linear regression.

https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

Data distribution and Model fitting







Fit a LR model
SBP ~ Age + BMI

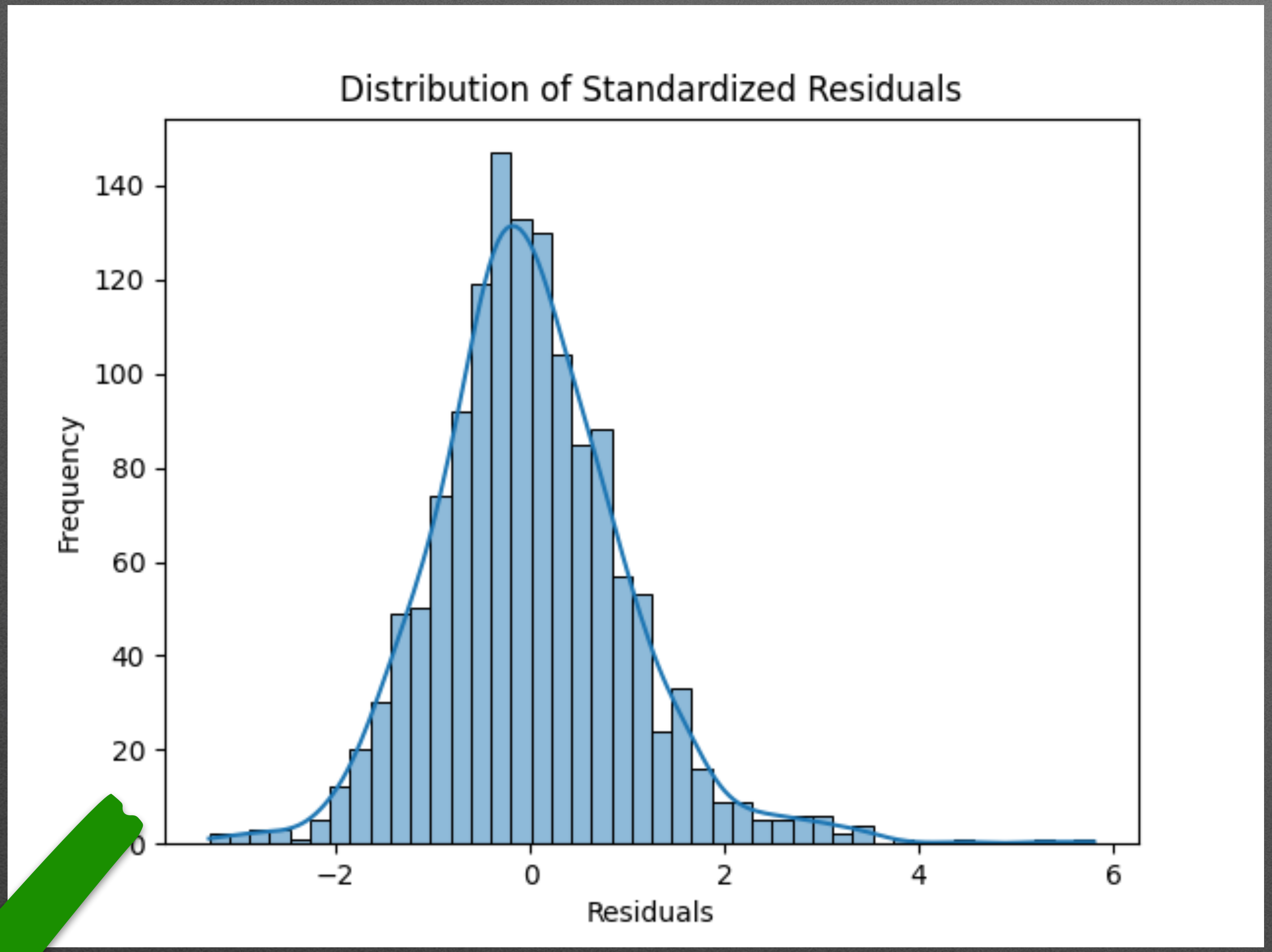
Coef Age: 0.44
Coef BMI: 0.31

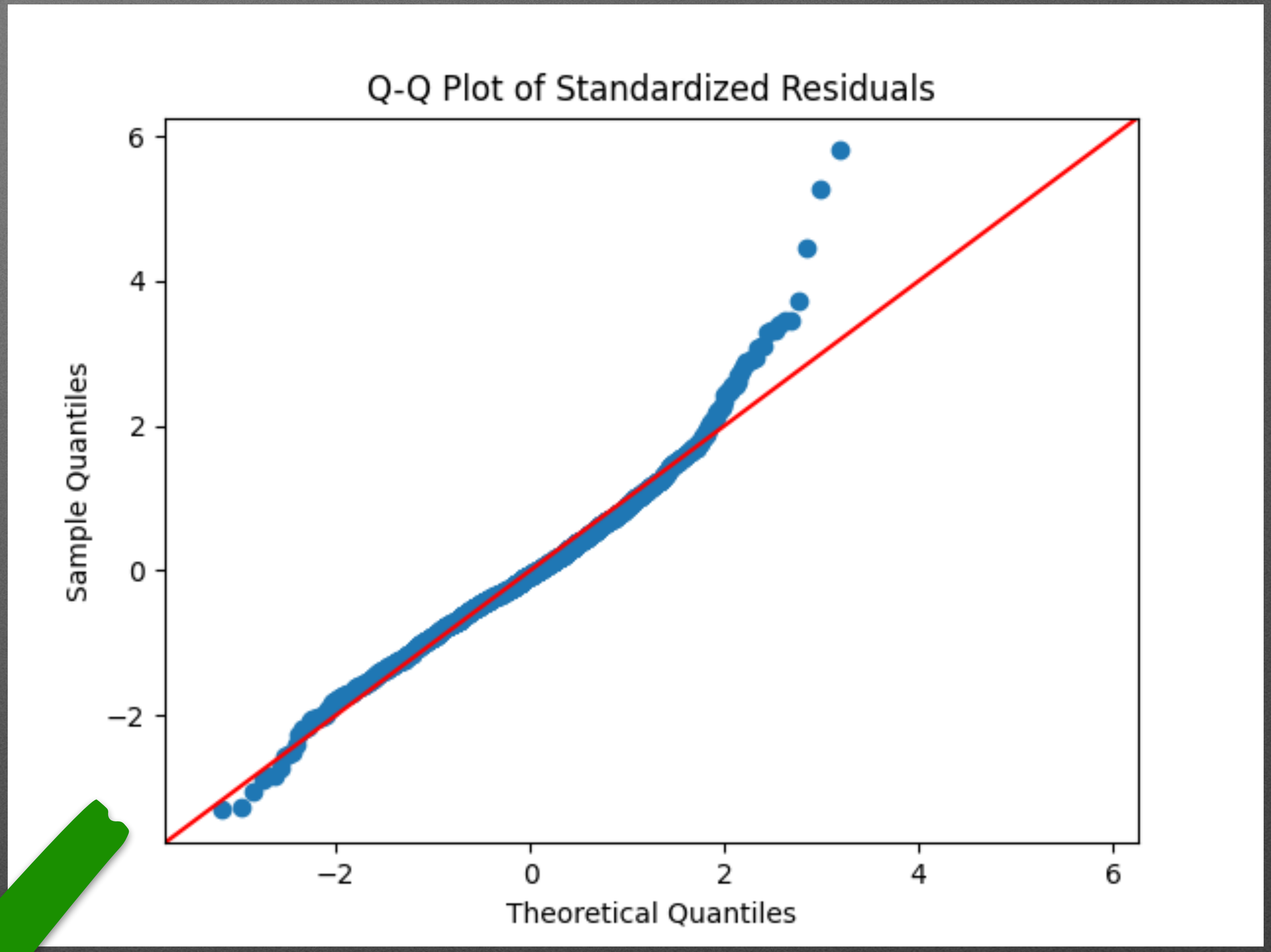
Does this relationship exist at the population level too?

Yes because
P-value < 0.0001

BUT…

Residual normality assumptions





What about the statistical tests? They are too sensitive so they fail here due to the outliers!

Do we need them here? Not really because large sample size and Central Limit Theorem comes to our rescue!

Lognormal: Everything can't be normal!

- Lognormal model: log(random variable) ~ Normal distribution
- Approx. 40000 cases, 1580 surgery types
- **<u>Question:</u>** Whether the surgery times are normally or lognormally distributed?
- Shapiro Wilk test for goodness of fit; one for each surgery type

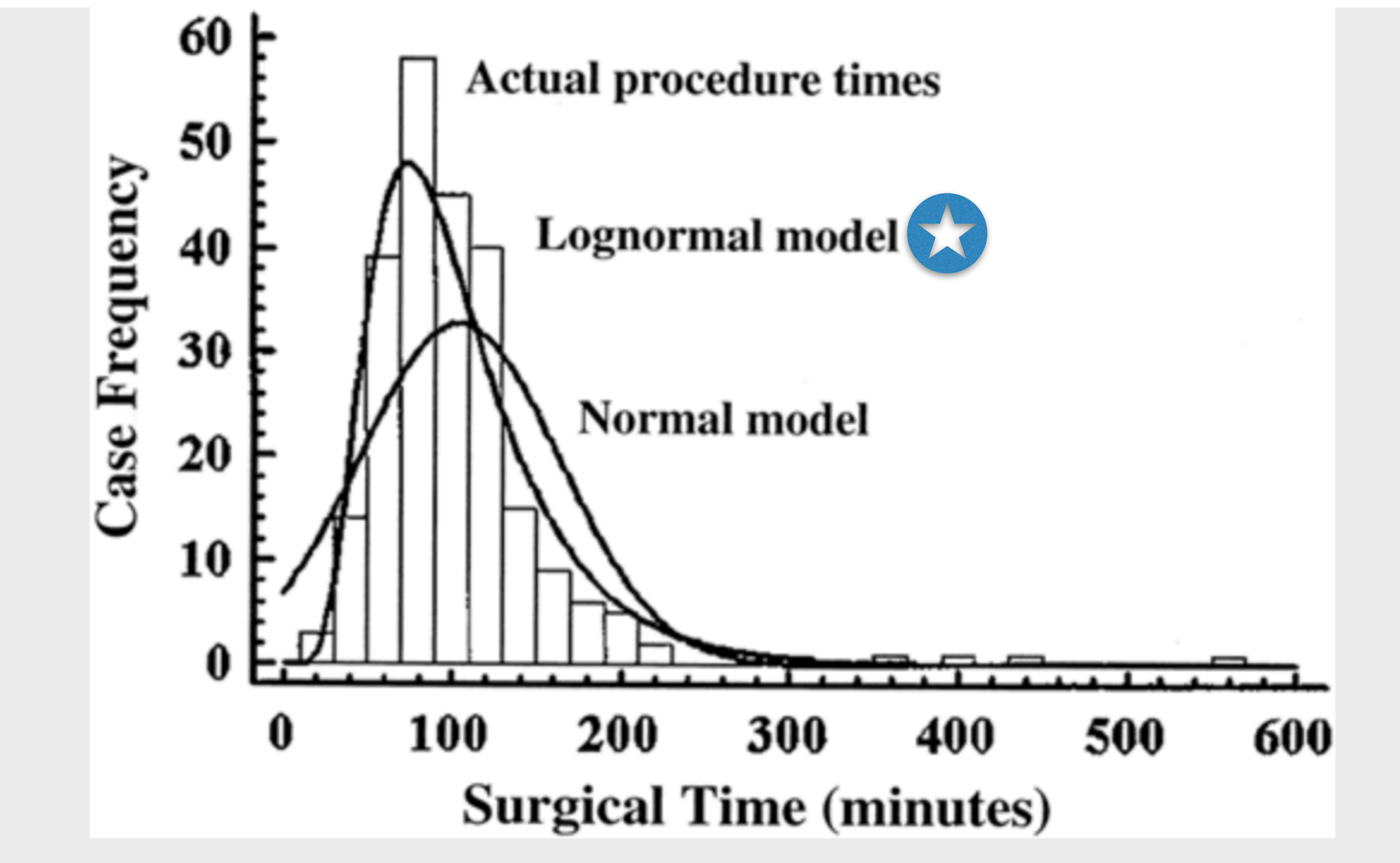


Fig. 1. To illustrate the differences between models, we compared the actual (empirical) data (illustrated as a frequency histogram) with the fitted ideal log-normal and normal models (the surgical procedure is exploration of the chest for postoperative bleeding during general anesthesia; n = 241 surgical cases).

Source: Strum, D.P., May, J.H. and Vargas, L.G., 2000. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *The Journal of the American Society of Anesthesiologists*, *92*(4), pp.1160-1167.

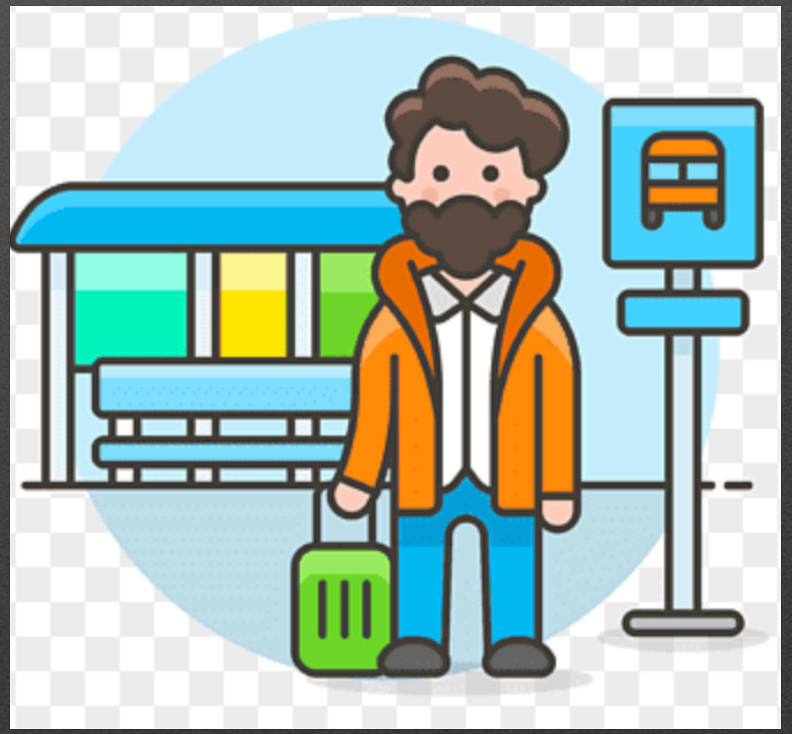
Exponential's memorylessness property

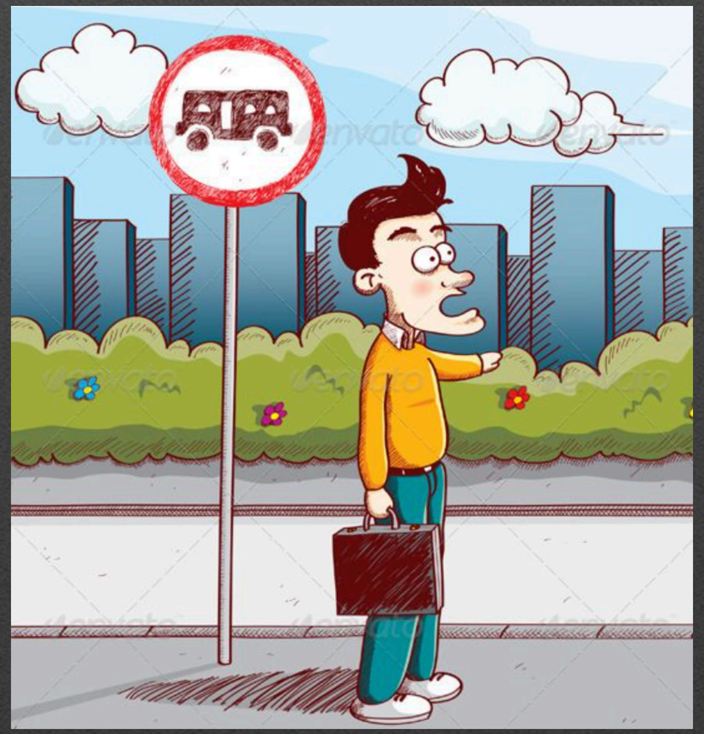
$X \sim Exponential(\lambda)$

$P(X > s + t \,|\, X > t) = e^{-\lambda s}$

Invariant of t !!

Can you imagine time between bus arrivals being exponential?





- Useful in reasoning average behavior in queueing systems
- If a stream of events have exponentially distributed inter-arrival times, then the expected time until the next event is always 1/λ , no matter how long one is waiting.


THANK YOU FOR LISTENING
DO YOU HAVE ANY QUESTIONS ?