STAT 515 fa 2023 Lec 07 slides

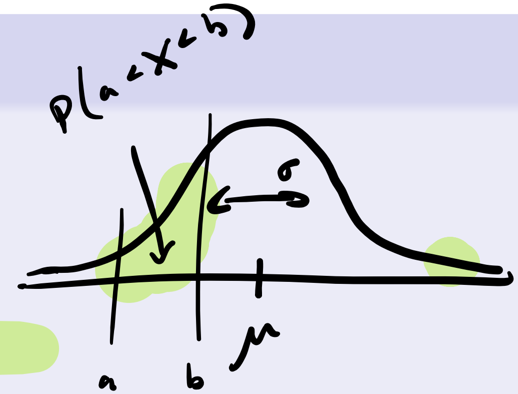
The Normal distribution

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Normal or Gaussian probability distribution



A continuous rv X with pdf given by

"mu"

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

sigma

has the *Normal distribution* with mean μ and variance σ^2

Carl Friedrich Gauss.

We write $X \sim \text{Normal}(\mu, \sigma^2)$.

The pdf of the Normal(μ, σ^2) distribution:

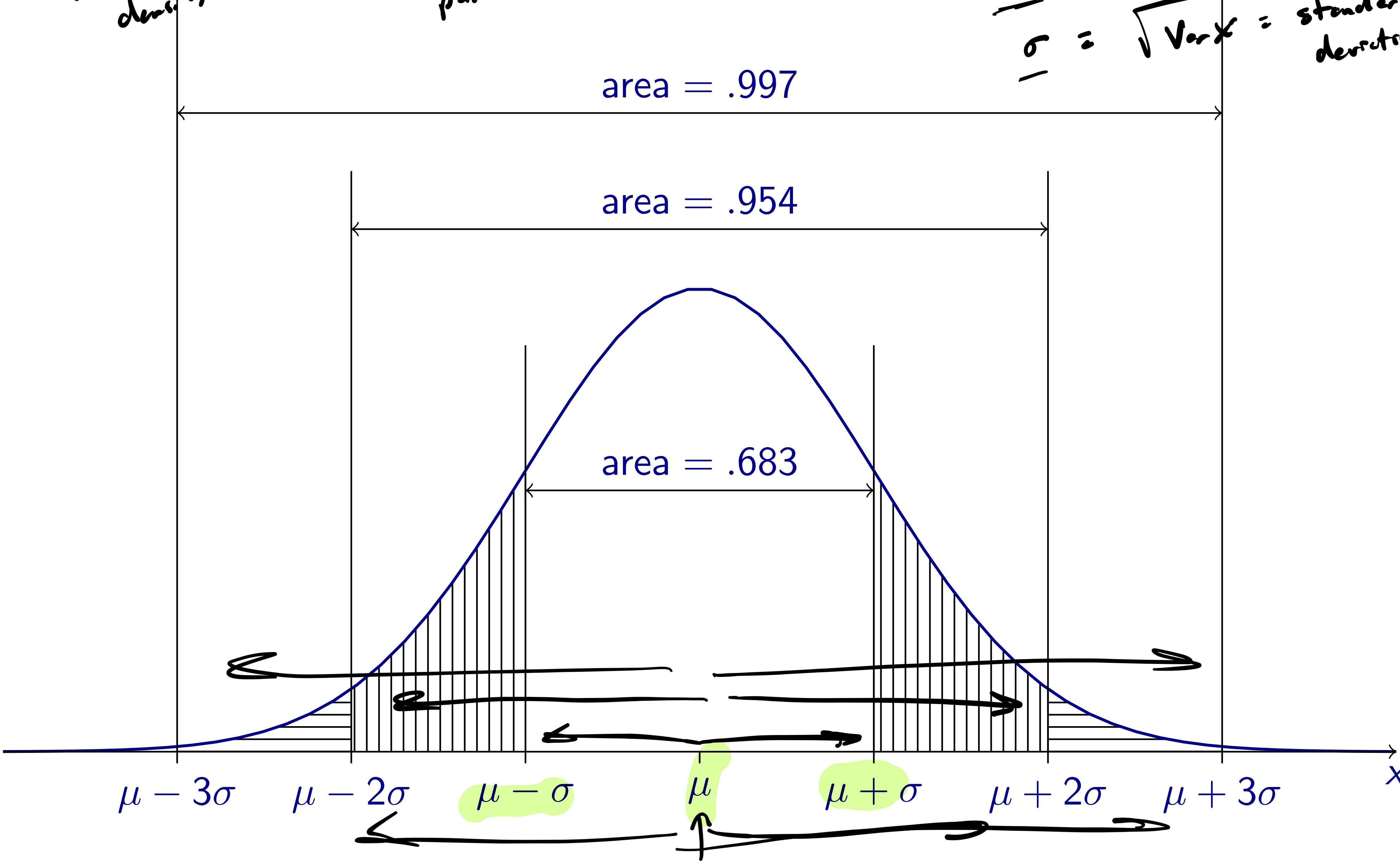
probability density function

parameters

$\mu = \mathbb{E}X$

$\sigma^2 = \mathrm{Var}\,X$

$\sigma = \sqrt{\mathrm{Var}\,X}$ = standard deviation



Mean and variance of Normal distribution

If $X \sim \text{Normal}(\mu, \sigma^2)$, then

- $\mathbb{E}X = \mu$
- $\text{Var}\, X = \sigma^2$, $\sigma = \sqrt{\text{Var}\, X}$ = standard dev.

Exercise: Suppose growth in height (ft) of Loblolly pines from age three to five is Normal($\mu = 5, \sigma^2 = 1/4$). Give the probability that the growth of a randomly selected Loblolly pine is $\sigma = \frac{1}{2}$

1. between 4.5 and 5.5 feet.
2. more than 7 feet.
3. less than 5.5 feet.
4. between 3.5 feet and 5.5 feet.

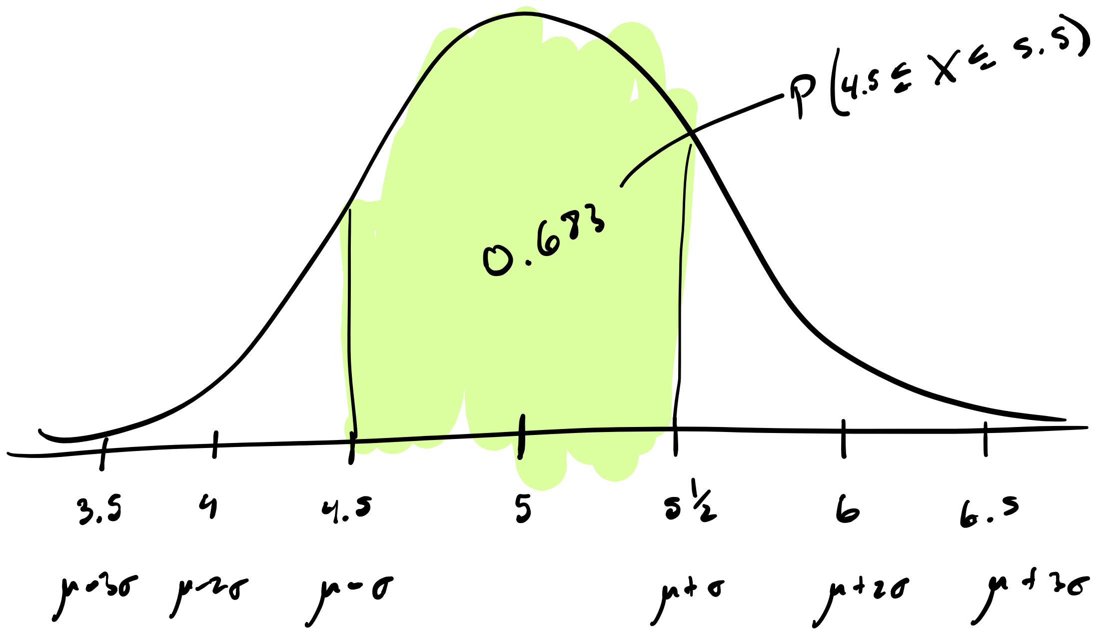
$(4.5, 5.5) = (\overset{5}{\mu} - \overset{1/2}{\sigma}, \mu + \sigma)$

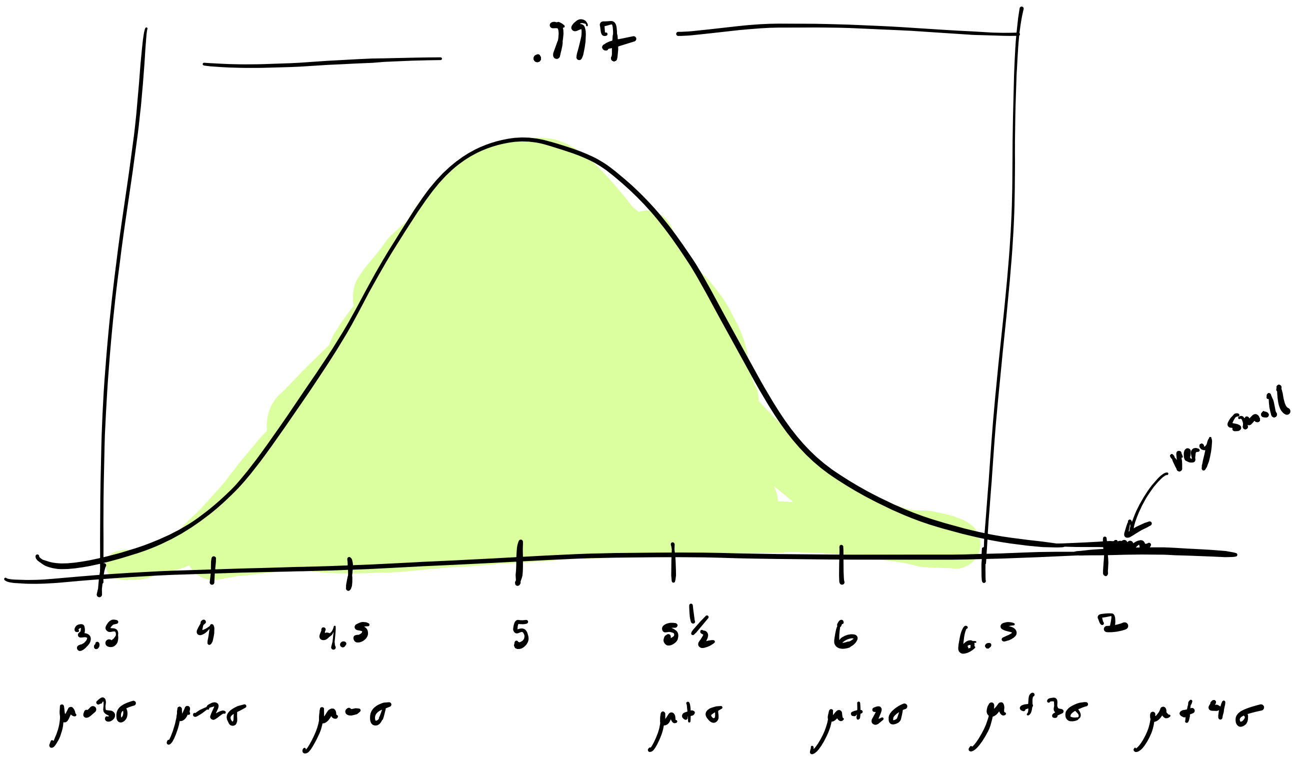
Use the picture on the previous slide.

① $P(4.5 < X < 5.5) = 0.683$

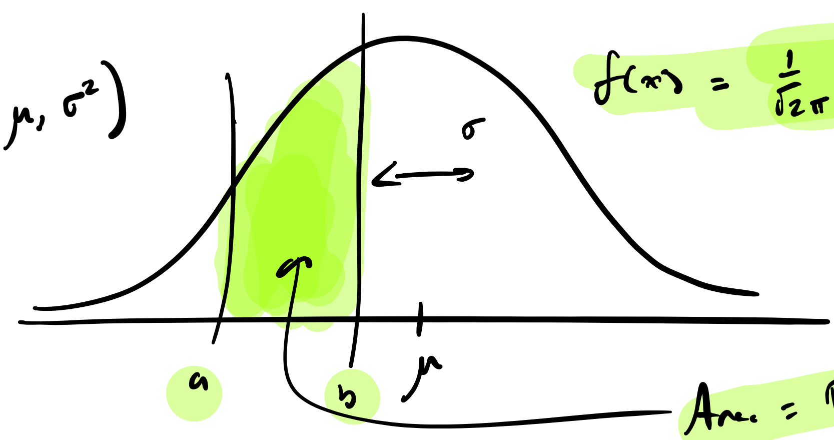
(2)

$\mu = 5$
$\sigma = \frac{1}{2}$





$N(\mu, \sigma^2)$



$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[- \frac{(x-\mu)^2}{2\sigma^2} \right]$$

$$\text{Area} = P(a < X < b) = \int_a^b \bigcirc \, dx$$

Get probabilities for $X \sim \text{Normal}(\mu, \sigma^2)$ like

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

Conversion to the Standard Normal distribution

If $X \sim \text{Normal}(\mu, \sigma^2)$, then

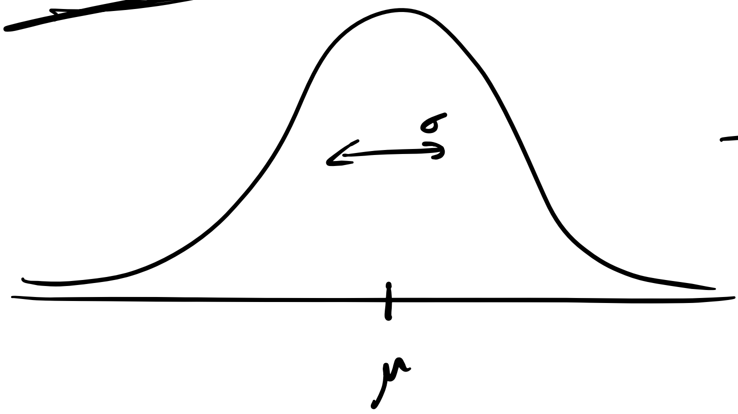
$$Z = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1).$$

The $\text{Normal}(0, 1)$ dist. is called the *Standard Normal distribution* and its pdf is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

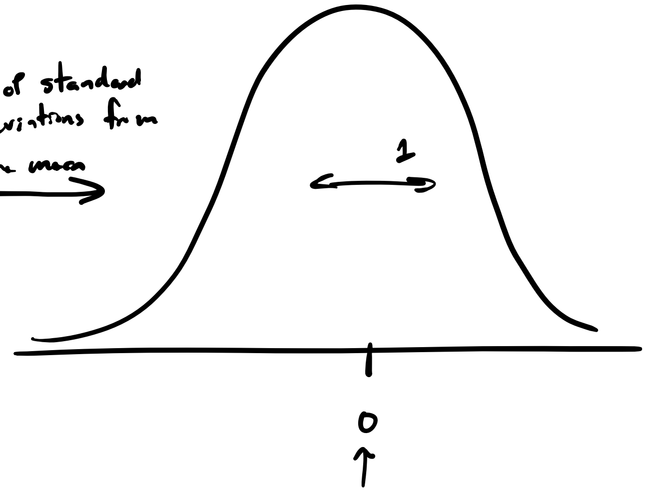
Can look up integrals over this pdf in a table.

$X \sim N(\mu, \sigma^2)$



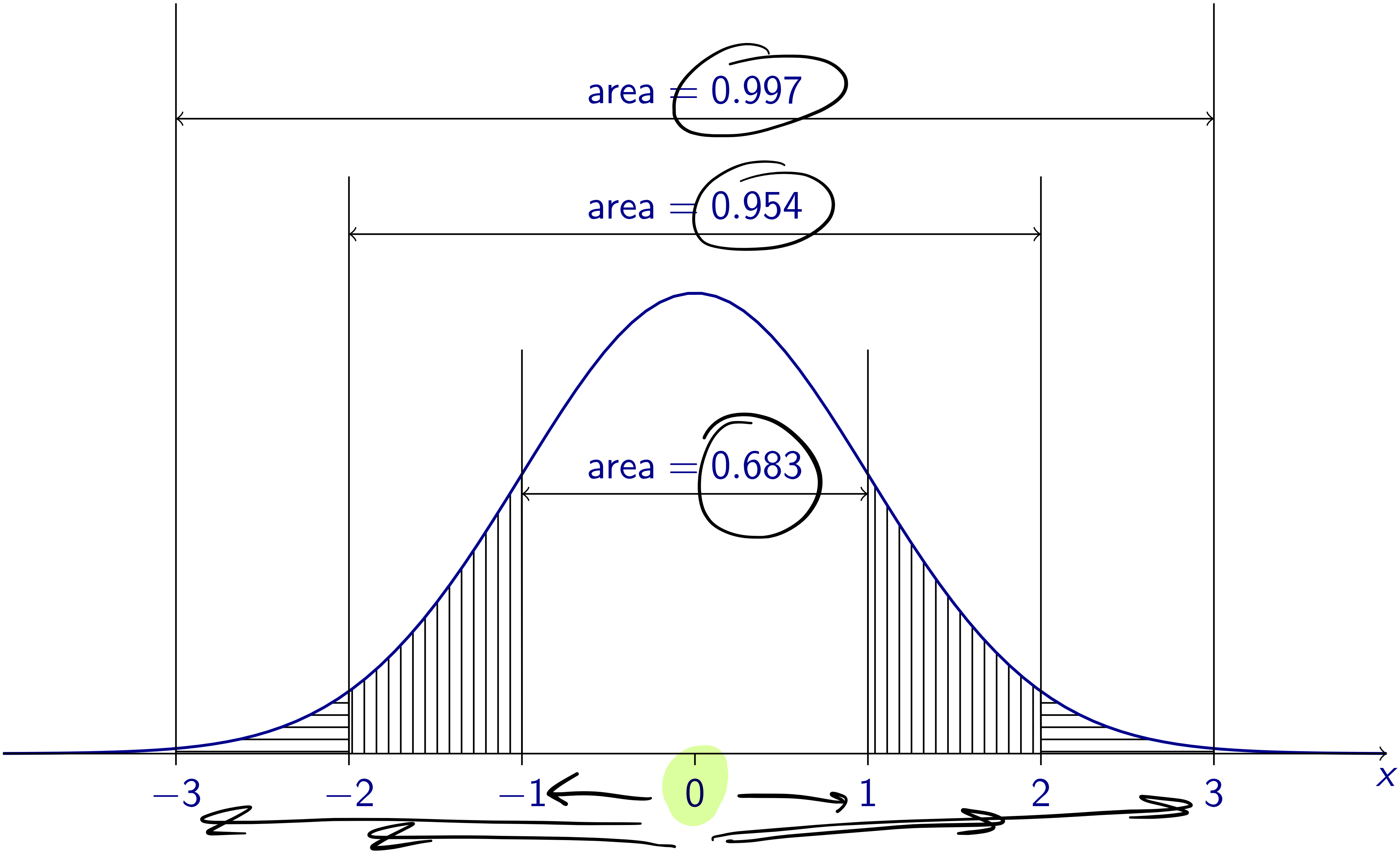
$Z = \frac{X-\mu}{\sigma} \sim N(\overset{\mu}{0}, \overset{\sigma^2 = 1}{1})$

= # of standard deviations from the mean

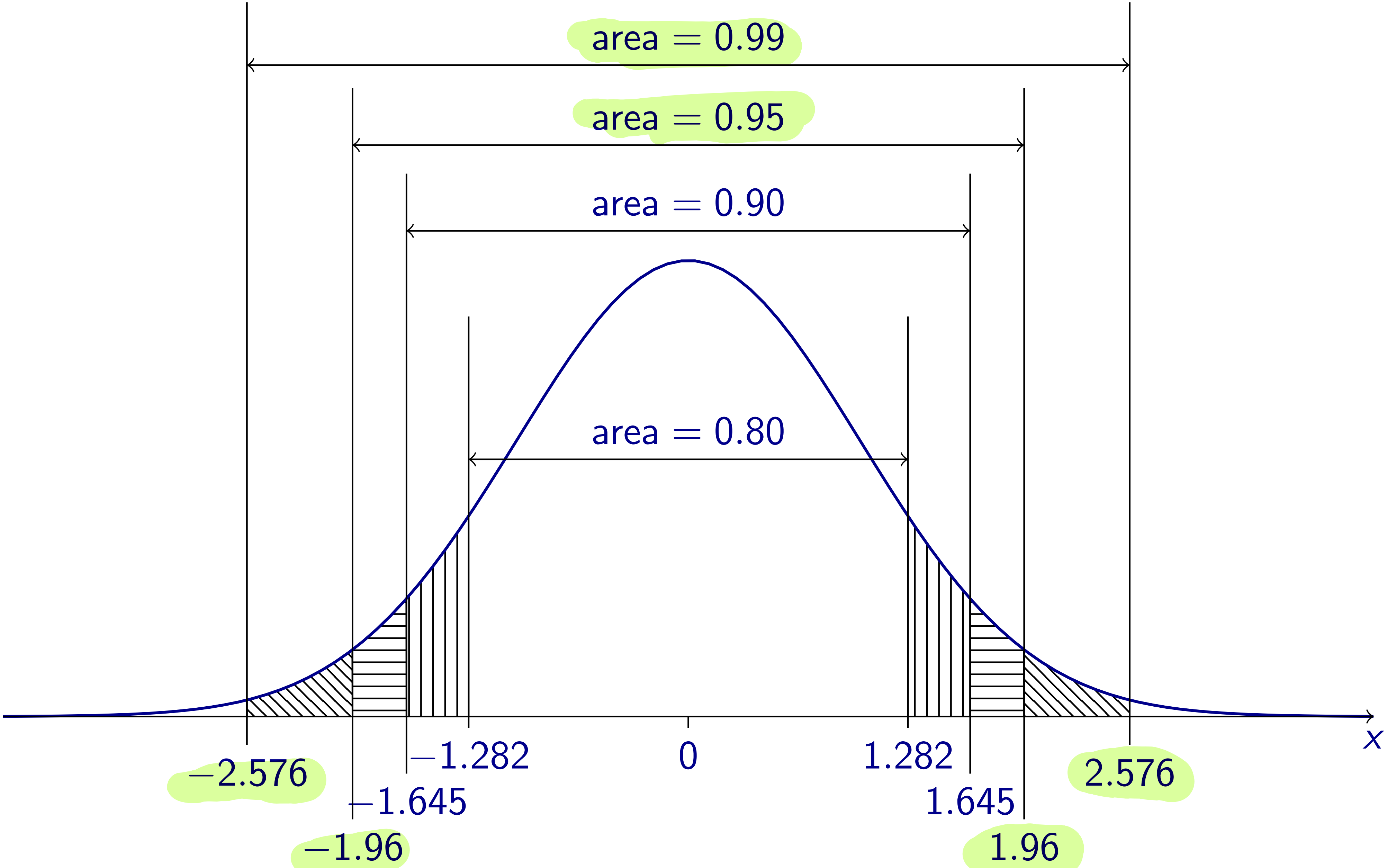


Standard Normal Distribution

The pdf of the Normal$(0,1)$ distribution:



The pdf of the Normal$(0, 1)$ distribution:



If $X \sim \text{Normal}(\mu, \sigma^2)$, we can find $P(a < X < b)$ in two steps:

1. Transform a and b to the Z-world (# of standard deviations world):

$$a \mapsto \frac{a - \mu}{\sigma} \quad \text{and} \quad b \mapsto \frac{b - \mu}{\sigma},$$

2. Find

$$P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$$

by using a "Z-table"—a table of Standard Normal probabilities.

Exercise: Suppose growth in height (ft) of Loblolly pines from age three to five is Normal($\mu = 5, \sigma^2 = 1/4$). Give the probability that the growth of a randomly selected Loblolly pine is

1. between 5.25 and 6.25 feet.
2. more than 7.8 feet.
3. less than 5.25 feet.
4. between 4.1 feet and 5.2 feet.

Find the probabilities in the "Z-world" using a Z-table.

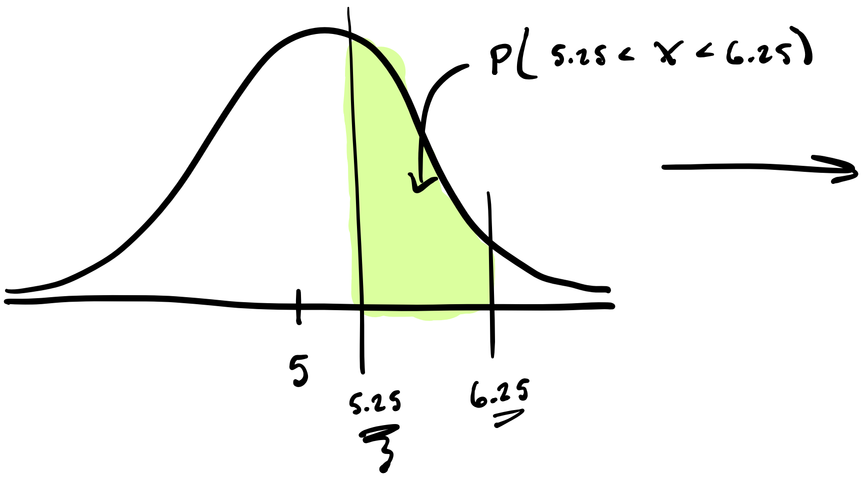
X = growth of randomly selected Loblolly pine tree from ages 3 - 5.

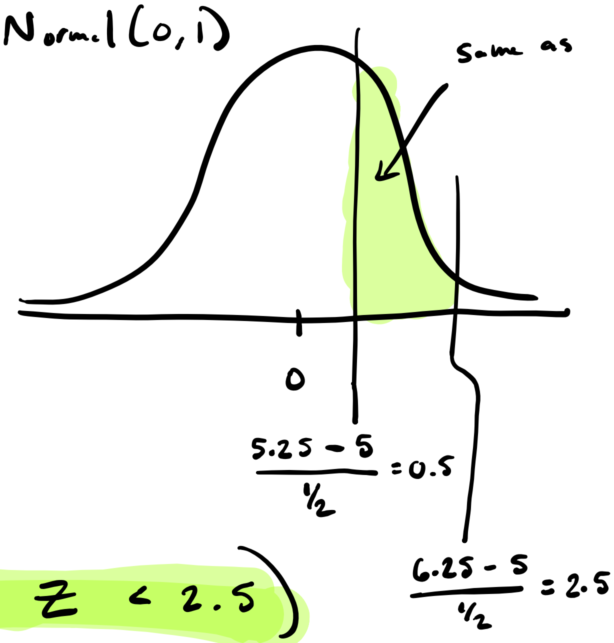
$X \sim \text{Normal}\left(\mu = 5, \sigma^2 = \frac{1}{4}\right), \quad \sigma = \frac{1}{2}$

$P(5.25 < X < 6.25)$

$X \sim \text{Normal}(\mu = 5, \sigma^2 = \frac{1}{4})$, $\sigma = \frac{1}{2}$

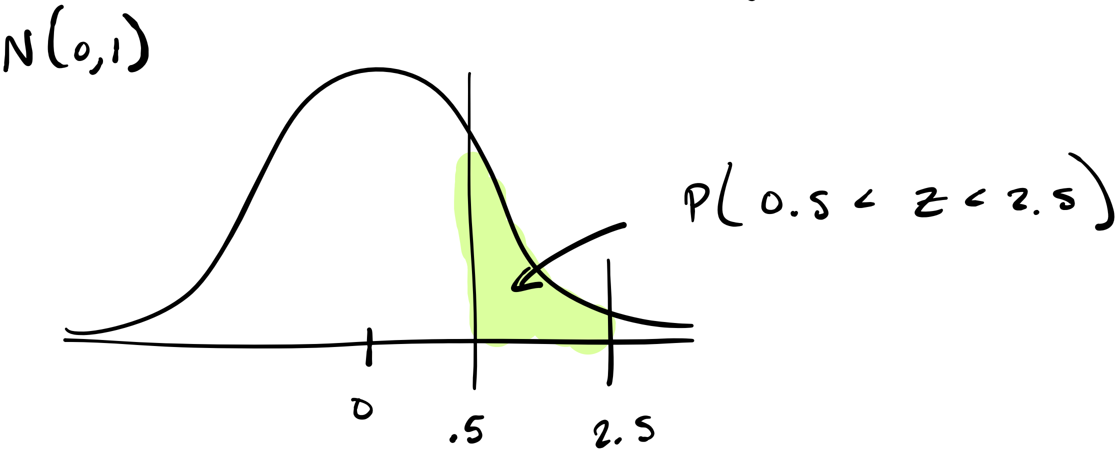
$Z = \frac{X - \mu}{\sigma}$ = # std. devs from μ

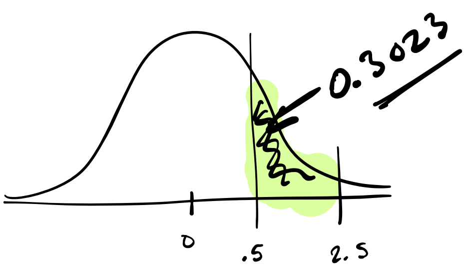




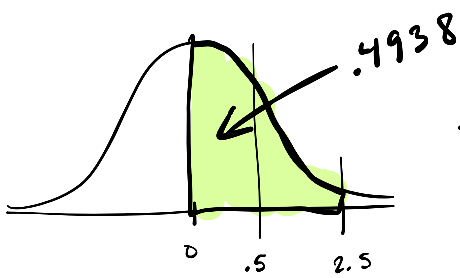
$P(5.25 < X < 6.25) = P(0.5 < Z < 2.5)$

Use table.

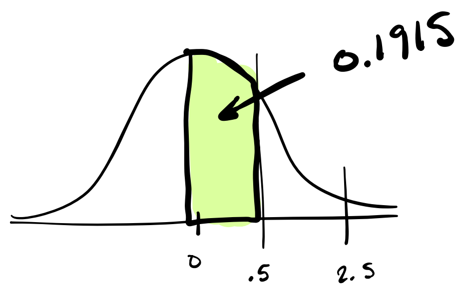




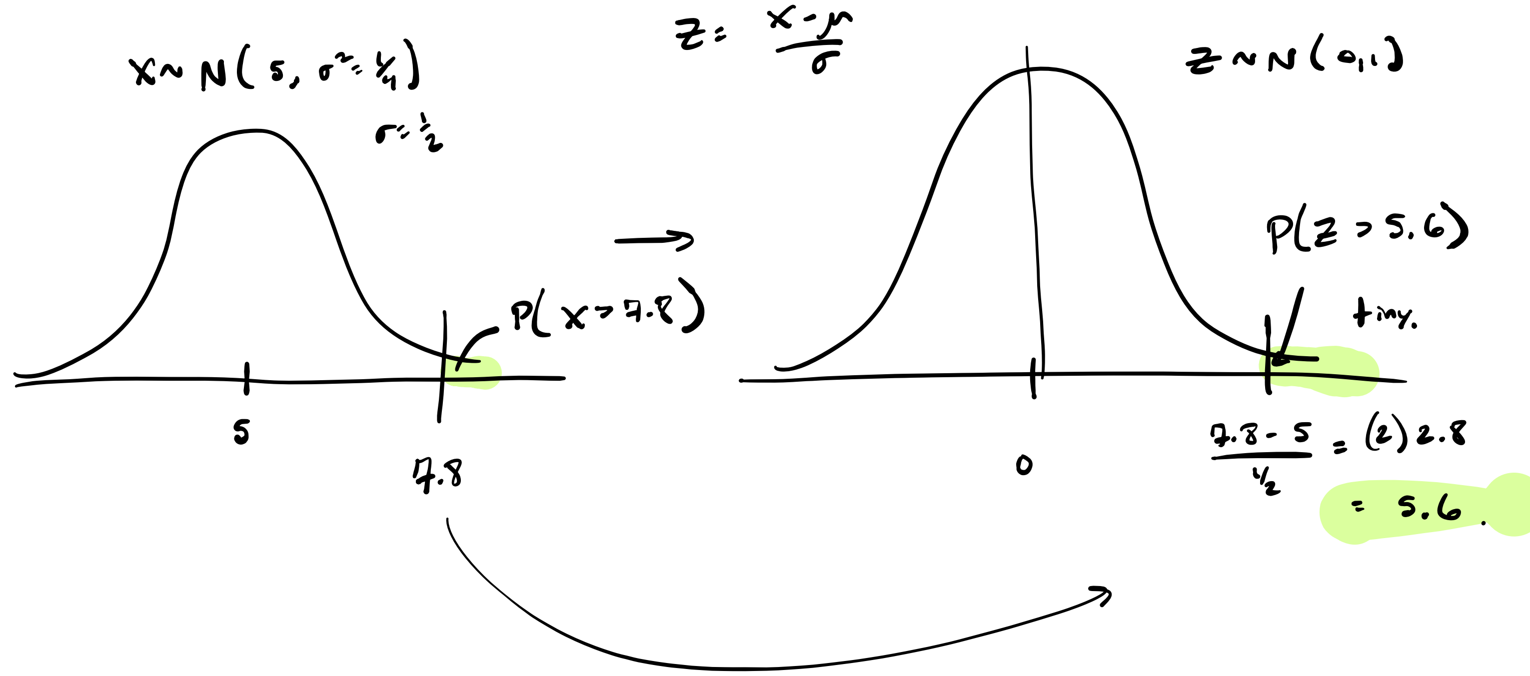
$=$



$-$

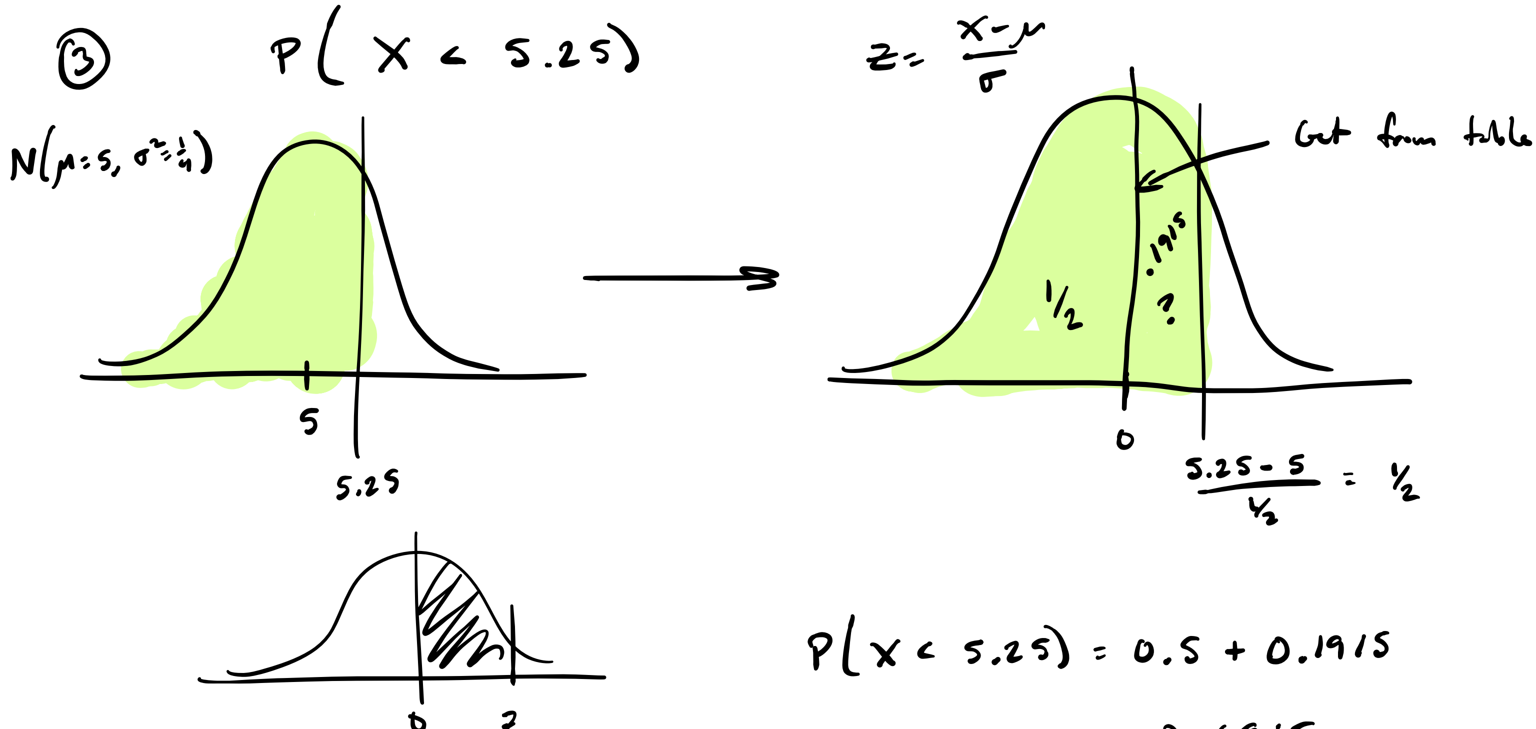


② $P(X > 7.8)$



Just say $P(X > 7.8) < \underline{0.001}$.

③ $P(X < 5.25)$



$P(X < 5.25) = 0.5 + 0.1915$

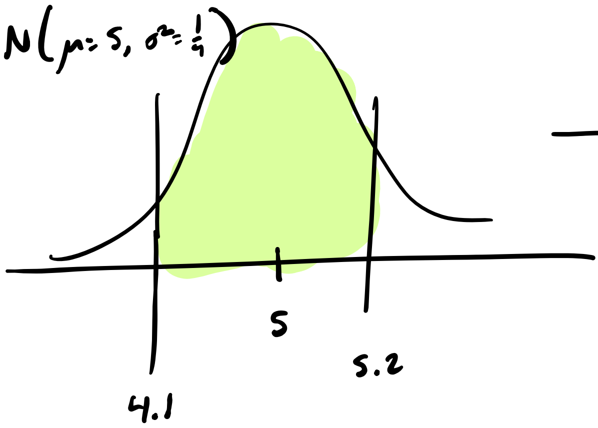
$= 0.6915.$

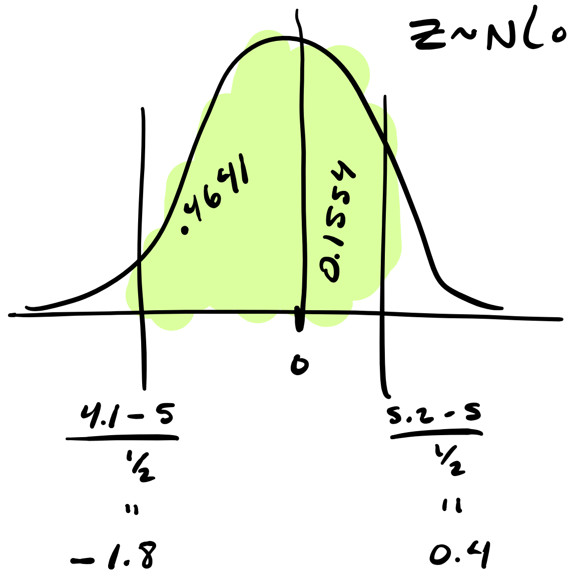
(4) $P(4.1 < x < 5.2)$

$z = \frac{x-\mu}{\sigma}$

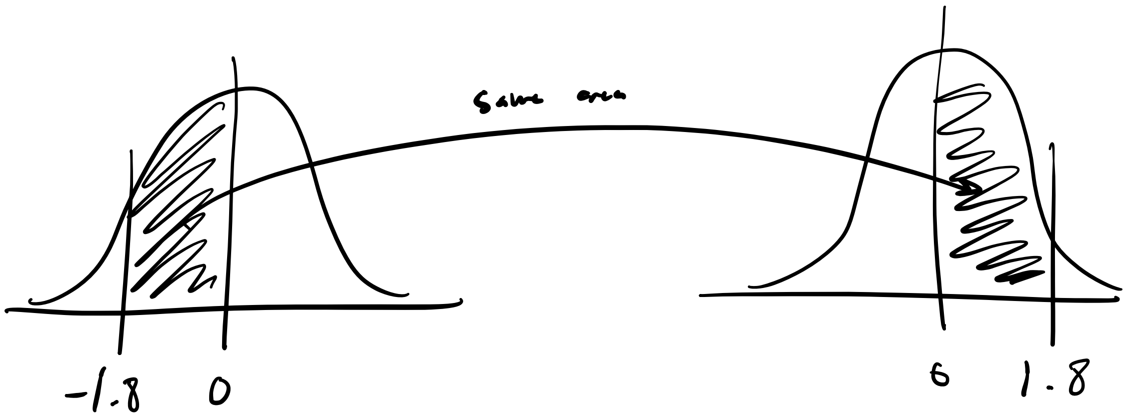
$X \sim N(\mu = 5, \sigma^2 = \frac{1}{4})$

$Z \sim N(0,1)$

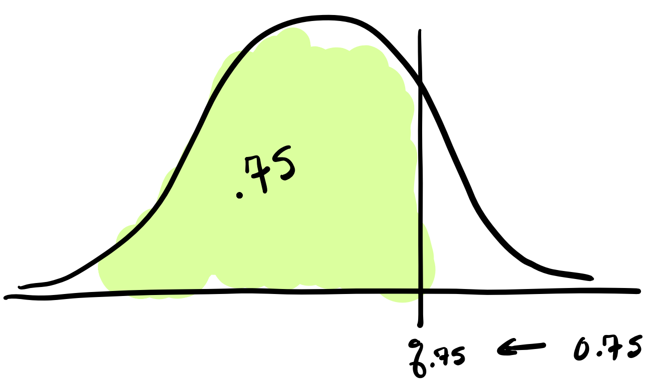




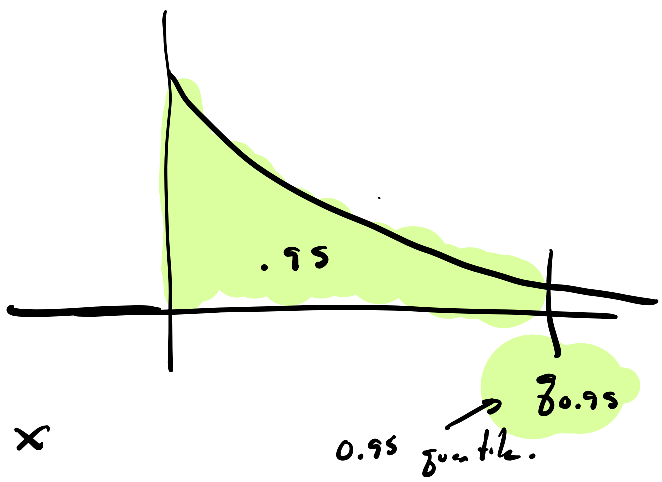
$P(4.1 < x < 5.2) = 0.4641 + 0.1554$



$X \sim$ Normal



← 0.75 quantile of X



θ = "theta"

Quantiles of a continuous random variable

For a continuous rv X with a strictly increasing cdf, the θth *quantile* of X is the value q_θ which satisfies

$$P(X \leq q_\theta) = \theta,$$

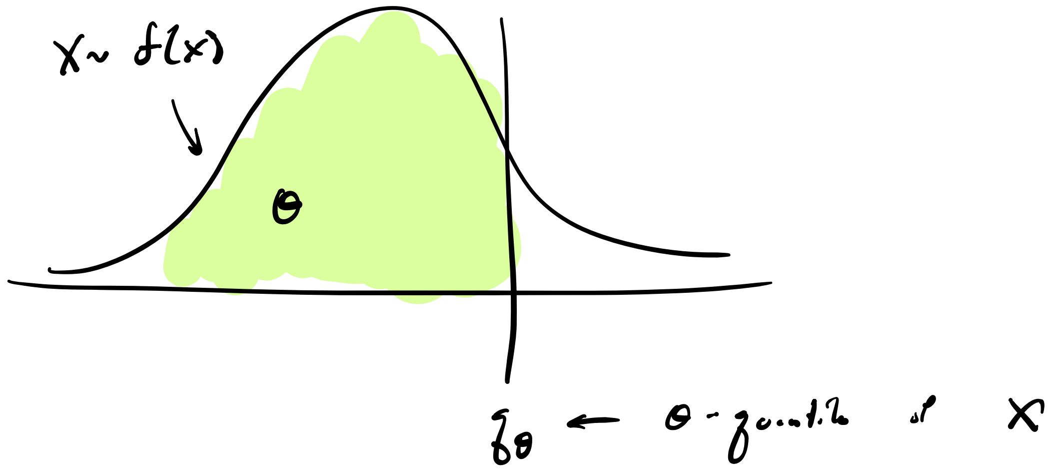
where θ is a value in $[0, 1]$.

Median is the 0.5 quantile

A quantile is like a percentile, but not expressed as a percentage.

Example: If X is the weight of a fresh chicken egg:

- With probability 0.90, a randomly selected egg has weight $\leq q_{0.90}$.
- With probability 0.25, a randomly selected egg has weight exceeding $q_{0.75}$.
- The median weight is $q_{0.50}$.


X ~ f(x)
θ
ξθ ← θ-quantile of X

If $X \sim \text{Normal}(\mu, \sigma^2)$, we can find q_θ such that $P(X \leq q_\theta) = \theta$ in two steps:

1. Find q_θ^Z such that $P(Z < q_\theta^Z) = \theta$ using a "Z-table".
2. Get the corresponding quantile in the X-world as

$$q_\theta = \mu + \sigma q_\theta^Z$$

Exercise: Suppose growth in height (ft) of Loblolly pines from age three to five is $\text{Normal}(\mu = 5, \sigma^2 = 1/4)$. Let X denote the height of a randomly selected Loblolly pine and find

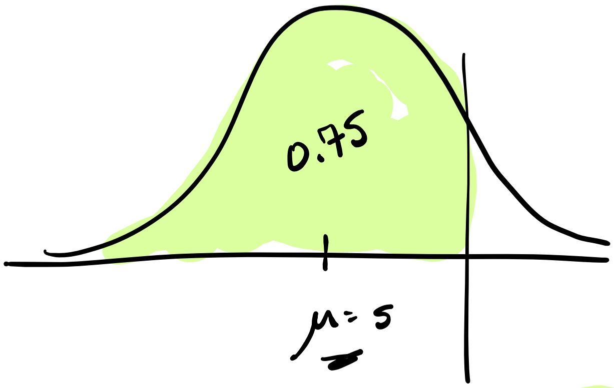
1. the 75%-tile of growth.
2. the median of the growths, i.e. the 50%-tile of X. because pdf is symmetric around $\mu = 5$, 5 is also the median.
3. an interval, centered at the mean, within which X lies with probability 0.50.

$X \sim \text{Normal}(\mu = 5, \sigma^2 = 1/4)$, $\sigma = 1/2$

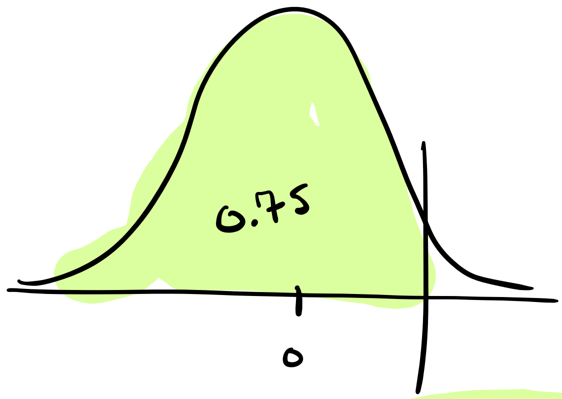
① Find 0.75 quantile of X

$$\boxed{Z = \frac{X-\mu}{\sigma} \iff X = \mu + \sigma Z}$$

$Z \sim N(0,1)$



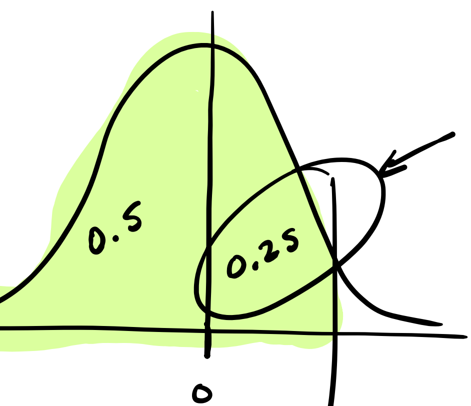
$q_{0.75} = ?$



$q^Z_{0.75} = 0.675$

Look up on table

$$q_{0.75} = 5 + \tfrac{1}{2}(0.675)$$
$$= 5 + .3375$$
$$= 5.3375$$

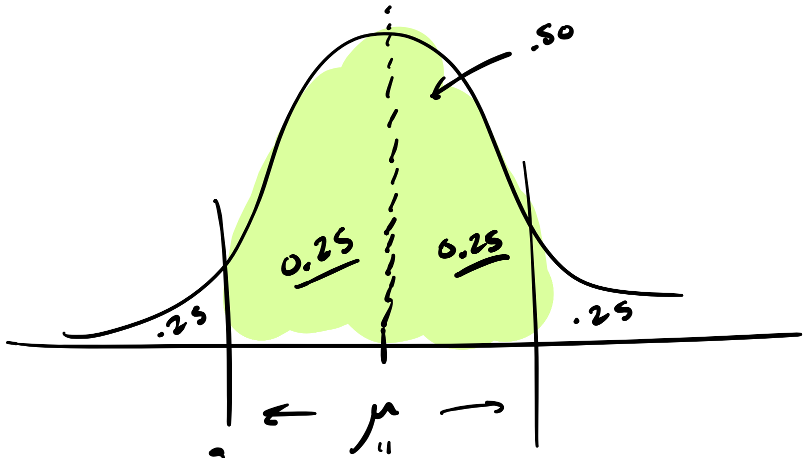


look in center of table for this value

margins of table

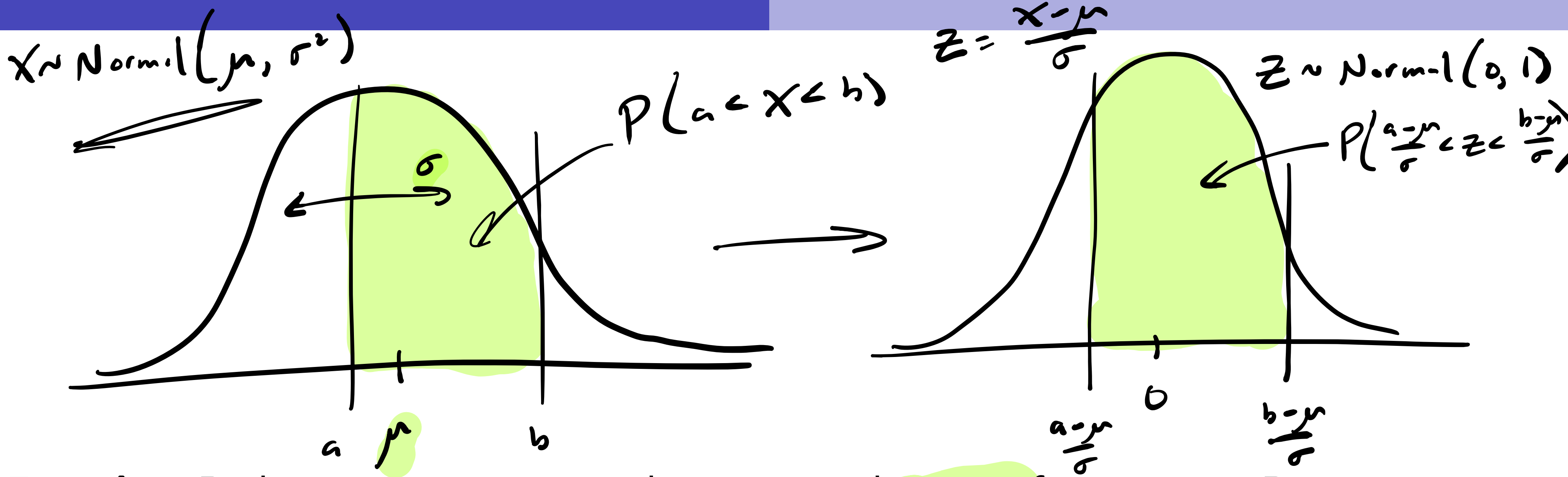
$q^Z_{0.75} = 0.675$

(3)



$z_{0.25}$ $\quad \mu = 5 \quad$ $z_{0.75} = 5.3375$

$z_{0.25} = 4.6625$



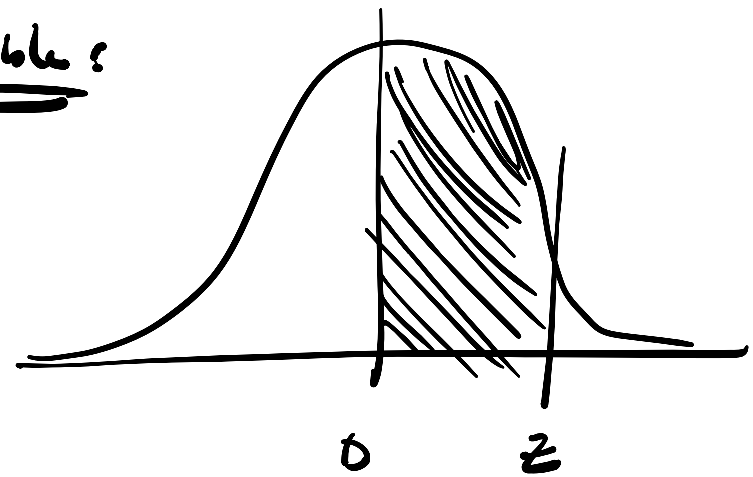
Exercise: Redo some exercises with `pnorm` and `qnorm` functions in R.

↑ probability ↑ quantile

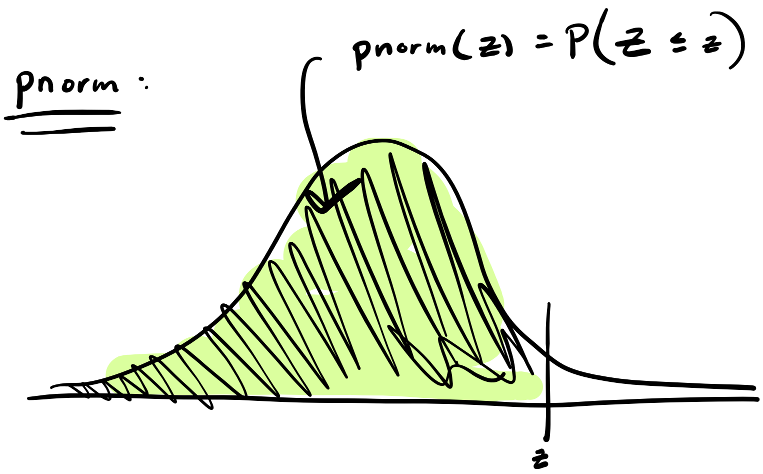
Exercise: Suppose growth in height (ft) of Loblolly pines from age three to five is Normal($\mu = 5, \sigma^2 = 1/4$). Give the probability that the growth of a randomly selected Loblolly pine is

z-table?

1. between 5.25 and 6.25 feet.
2. more than 7.8 feet.
3. less than 5.25 feet.
4. between 4.1 feet and 5.2 feet.



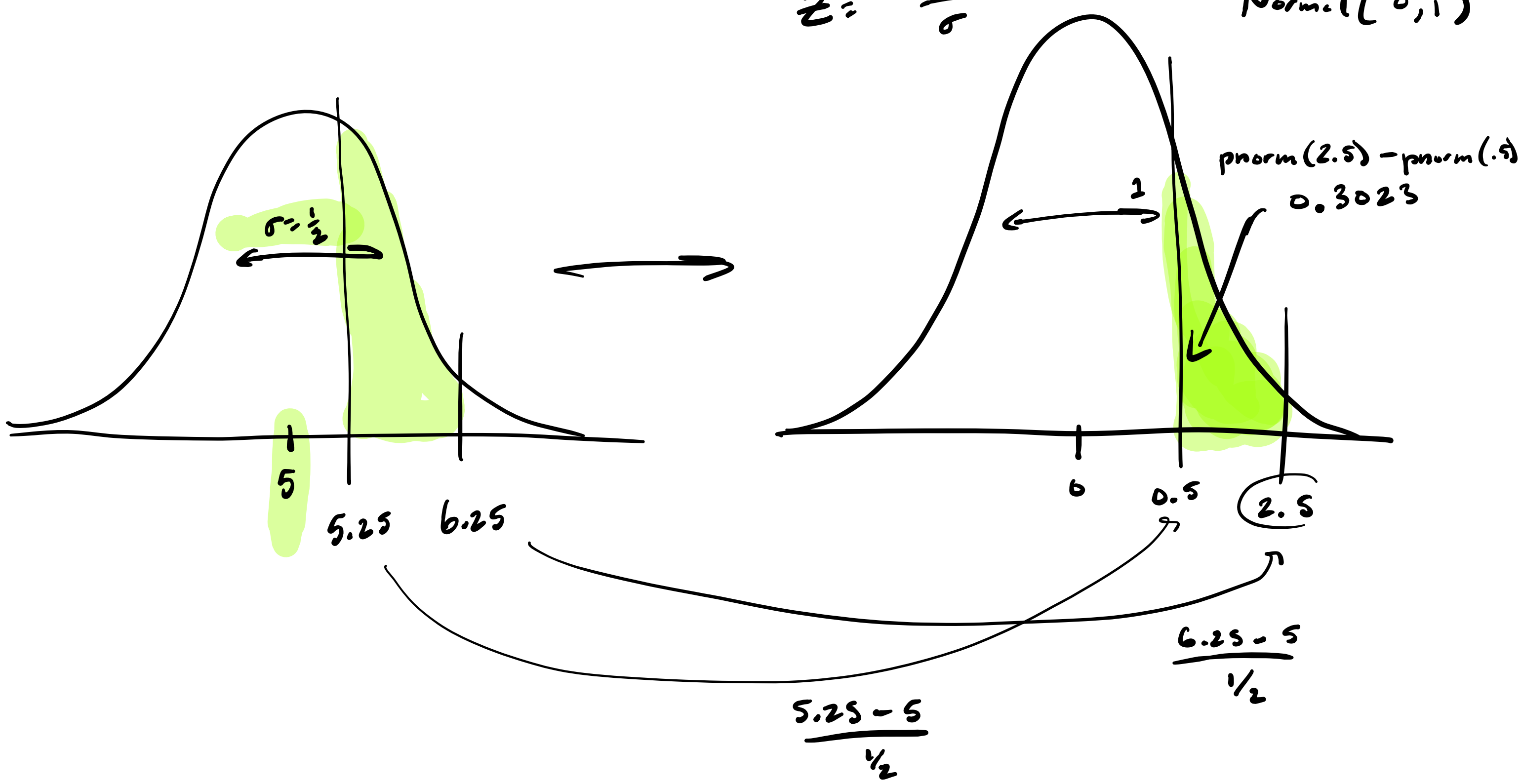
pnorm :



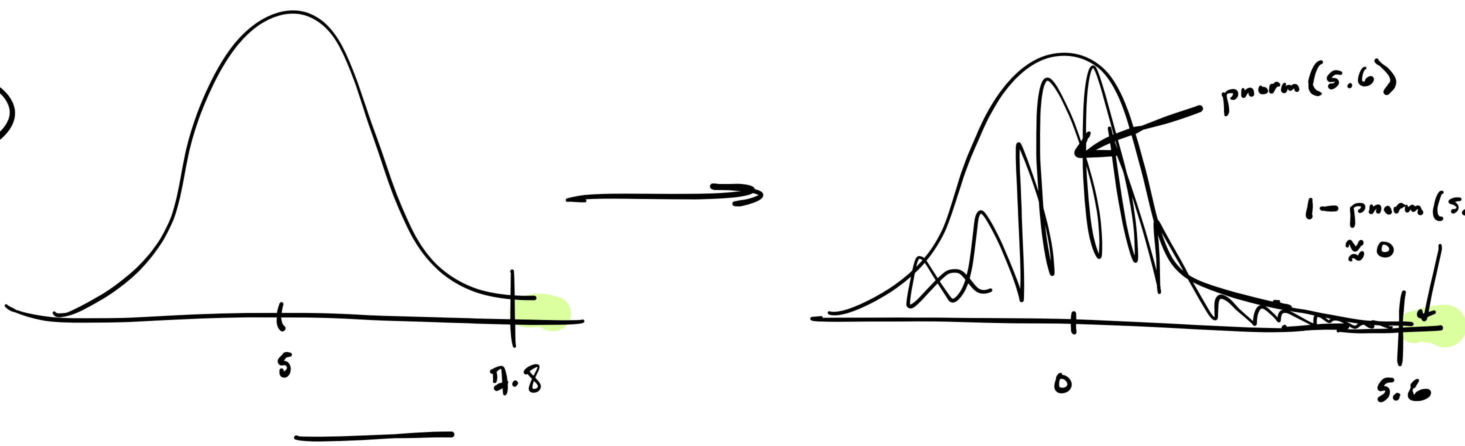
$X \sim \text{Normal}\left(\mu = 5, \ \sigma^2 = \frac{1}{4}\right), \ \sigma = \frac{1}{2}$

① $P(5.25 < X < 6.25)$

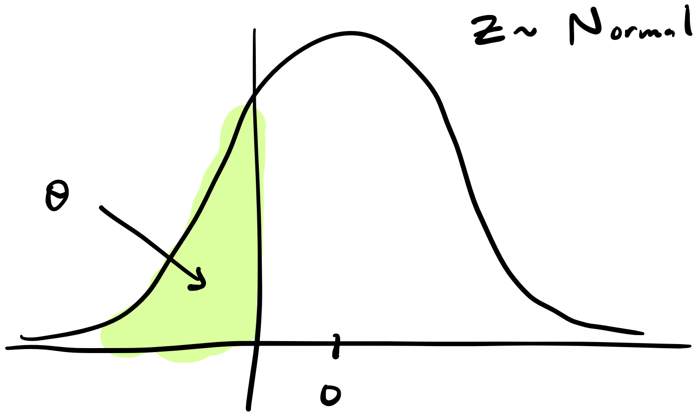
$Z = \frac{X - \mu}{\sigma}$



②



$Z \sim \text{Normal}(0,1)$



$q^Z_\theta = \theta^{th}$ quantile of Z = qnorm(θ)

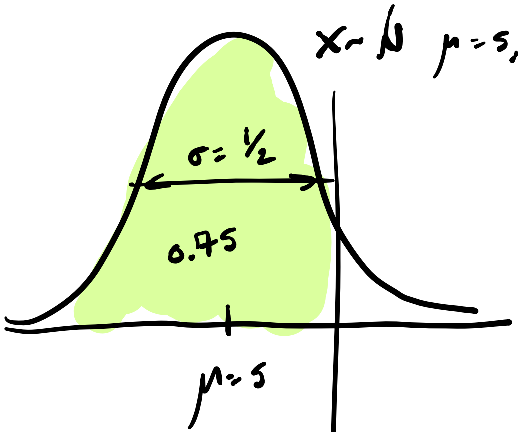
Exercise: Suppose growth in height (ft) of Loblolly pines from age three to five is Normal($\mu = 5, \sigma^2 = 1/4$). Let X denote the height of a randomly selected Loblolly pine and find

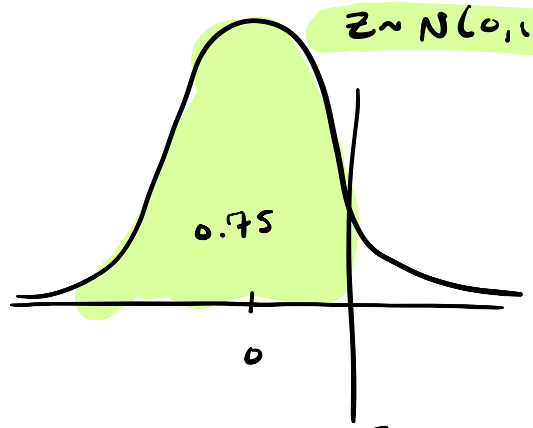
1. the 75%-tile of growth.
2. the median of the growths, i.e. the 50%-tile of X. ← because pdf is symmetric around $\mu = 5$, 5 is also the median.
3. an interval, centered at the mean, within which X lies with probability 0.50.

$Z = \frac{X-\mu}{\sigma} \iff X = \sigma Z + \mu$

$X \sim N(\mu = 5, \sigma^2 = 1/4)$

$Z \sim N(0,1)$





$q_{0.75} = \left(\frac{1}{2}\right)(0.674) + 5$

$= 5.3375$

$q^Z_{0.75} = \text{qnorm}(0.75)$

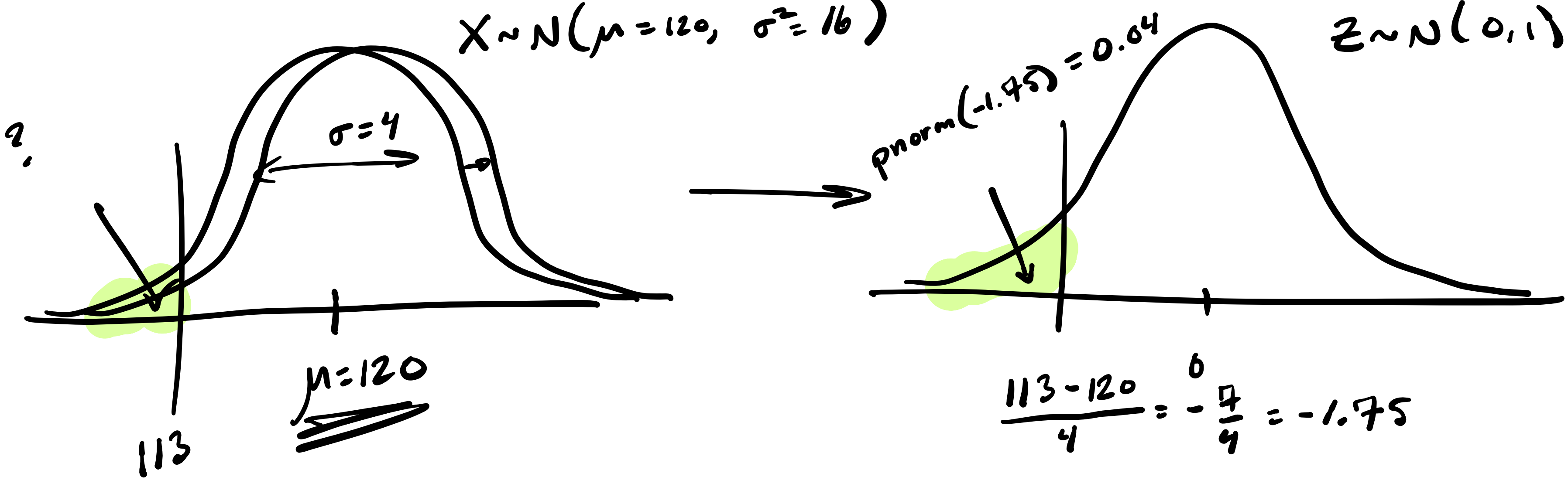
$= 0.674$

qnorm(0.5) = 0

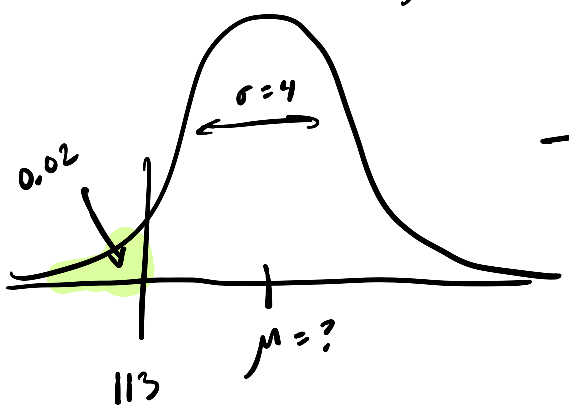
$4oz = 113g$

Exercise: You sell jars of baby food labelled as weighing 4oz ≈ 113g. Suppose your process results in jar weights with the $\boxed{\text{Normal}(\mu = 120, \sigma^2 = 4^2)}$ distribution and that you will be fined if more than 2% of your jars weigh less than 113g.

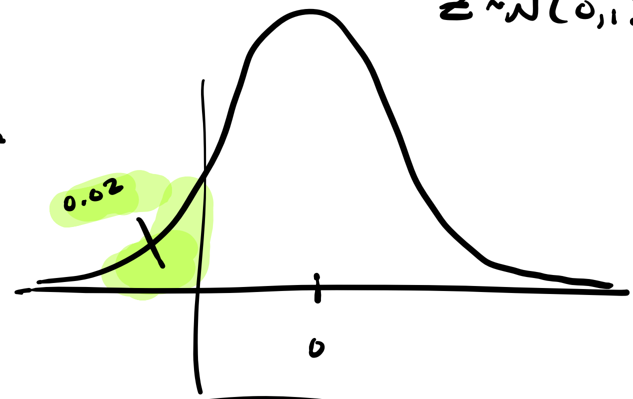
1. What proportion of your jars weigh less than 113g?
2. To what must you increase μ to avoid being fined?
3. Keeping $\mu = 120$g, to what must you reduce σ to avoid being fined?



② $X \sim N(\mu = ?, \sigma^2 = 16)$ $\qquad Z = \frac{X-\mu}{\sigma}$

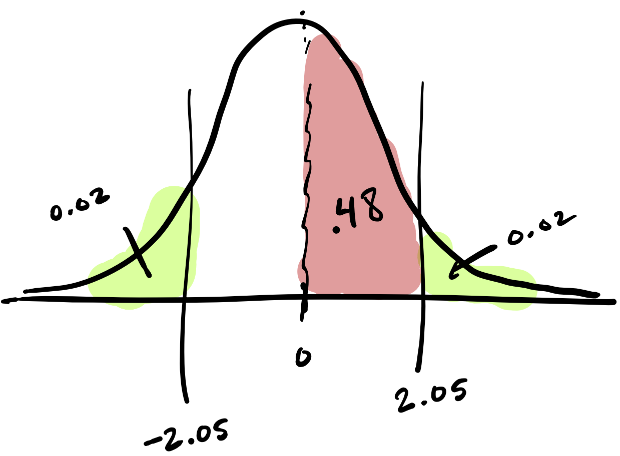


$Z \sim N(0,1)$



$$\boxed{\frac{113-\mu}{4} = \text{qnorm}(0.02) = -2.05}$$

Z-table way: $N(0,1)$



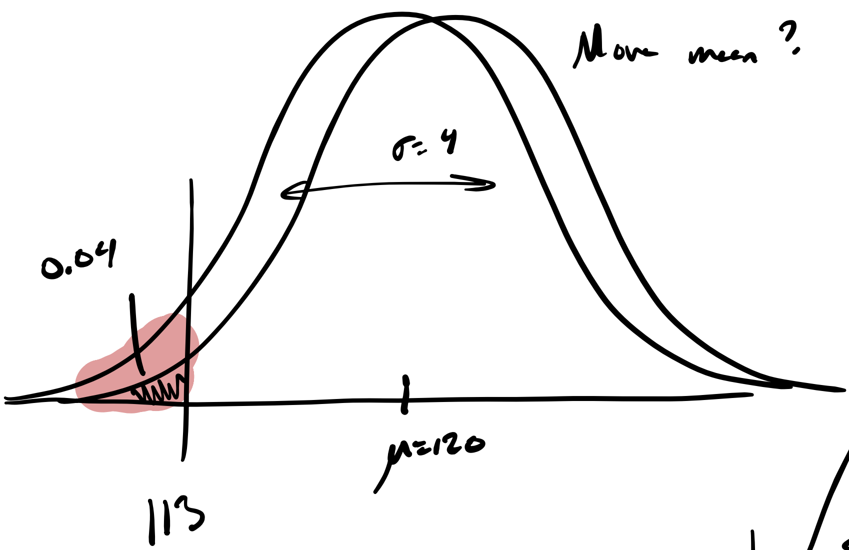
$$\frac{113-\mu}{4} = -2.05$$

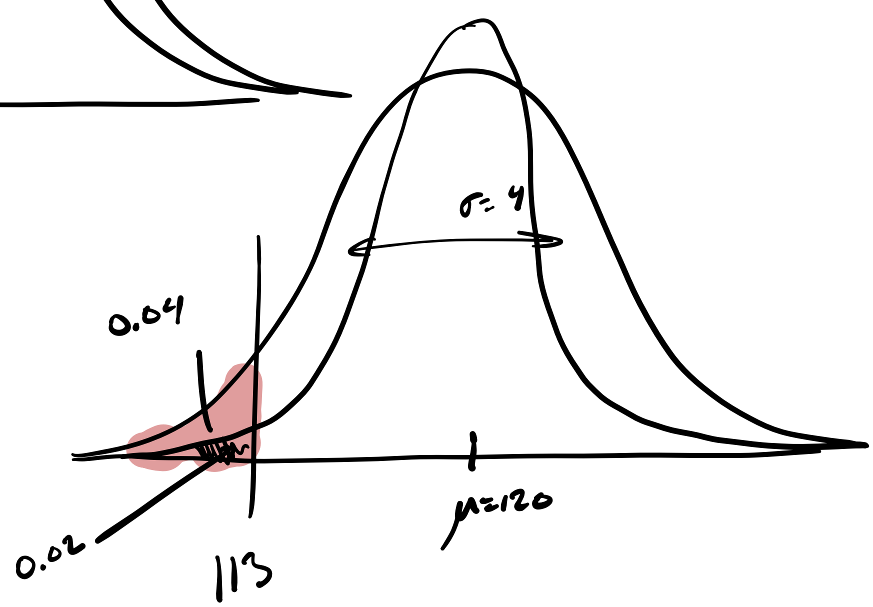
$$\Leftrightarrow \quad 113 - \mu = -4(2.05) = -8.2$$

$$\Leftrightarrow \quad 113 + 8.2 = \mu$$

$$\Leftrightarrow \quad \boxed{\mu = 121.2}$$

More mean?

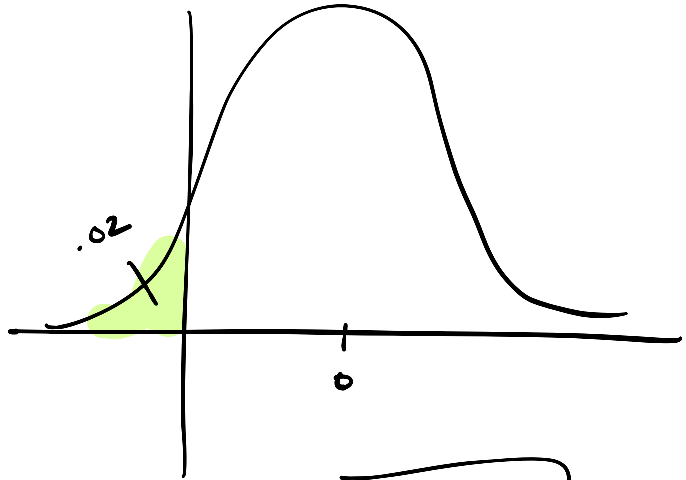




$X \sim N(\mu = 120,\ \sigma = ?)$

$z = \frac{x-\mu}{\sigma}$

$Z \sim N(0,1).$

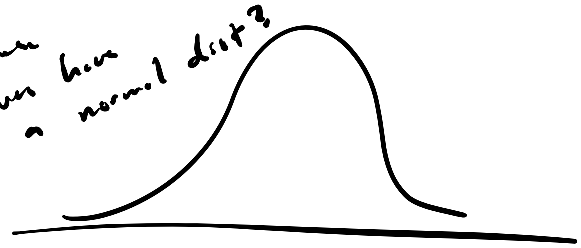


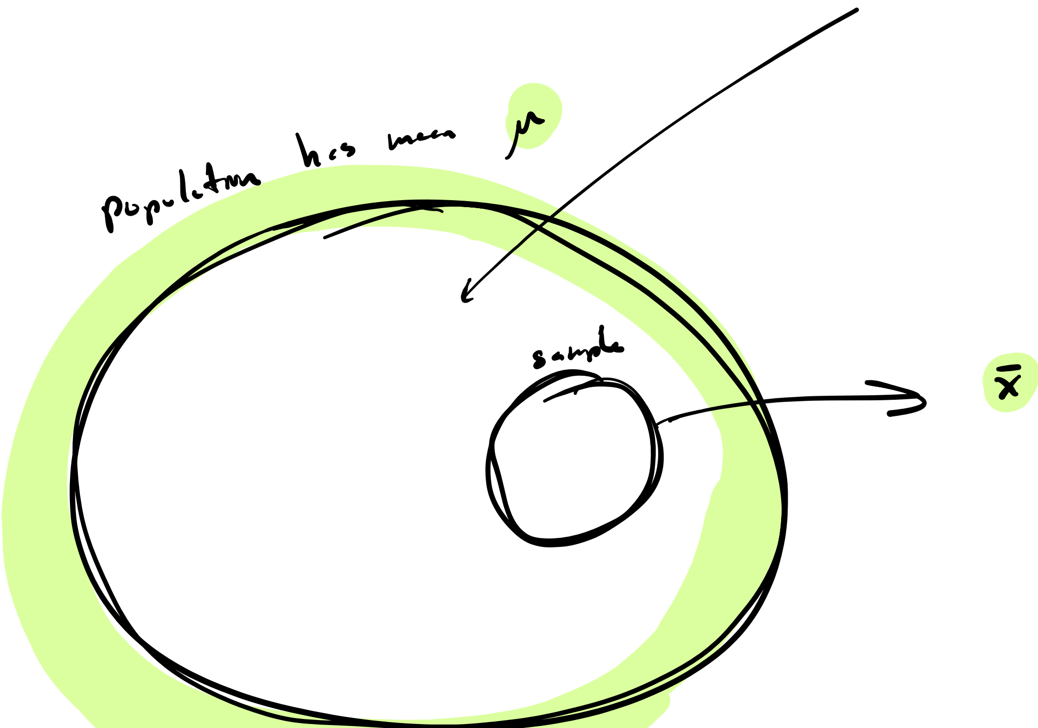
$$\frac{113-120}{\sigma} = \boxed{-2.05}$$

$$\iff \frac{-7}{-2.05} = \sigma$$

$$\iff \sigma = \frac{7}{2.05} \boxed{= 3.41}$$

Do these values have a normal dist?





Do my data come from a Normal distribution?

🤔

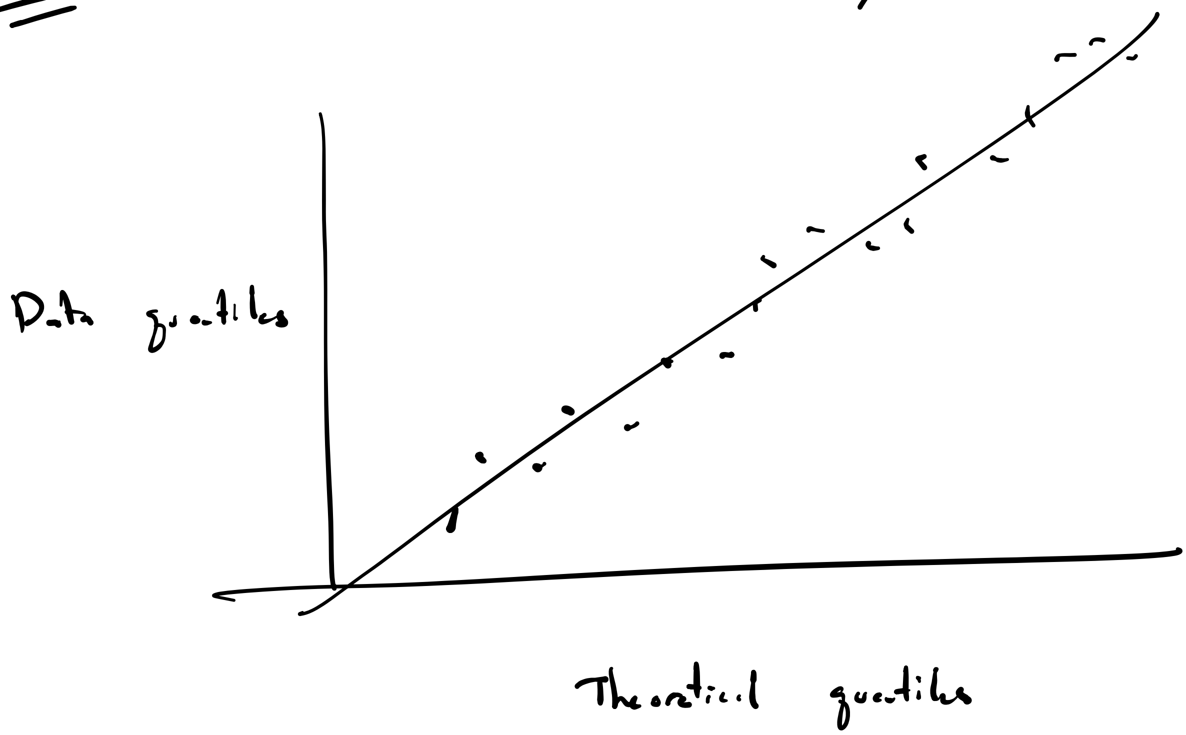
Example: These are the commute times (sec) to class of a sample of students.

1832	1440	1620	1362	577	934	928	998	1062	900
1380	913	654	878	172	773	1171	1574	900	900

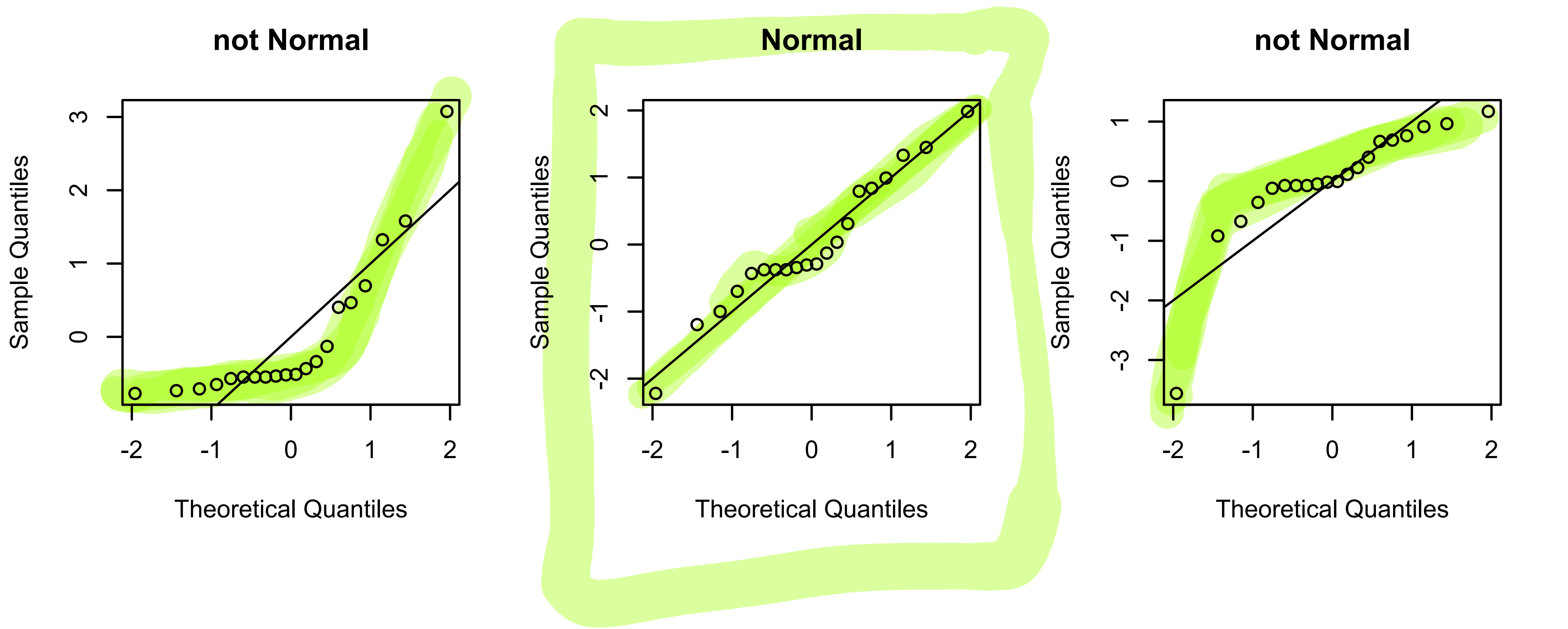
Check with a Q-Q plot whether the data quantiles match those of a Normal distribution.

QQ plot

Q: Are the data Normally distributed?



Some more Q-Q plots:



If $X_1, \dots, X_n \overset{ind}{\sim} N(\mu, \sigma^2)$ then $Y = X_1 + \dots + X_n \sim N(n\mu, n\sigma^2)$

Sum of independent Normal random variables

If $X_1 \sim \text{Normal}(\mu_1, \sigma_1^2), \dots, X_n \sim \text{Normal}(\mu_1, \sigma_n^2)$ are independent random variables, then

$$Y = \sum_{i=1}^{n} X_i \sim \text{Normal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

In the above, *independent* means that the values of the rvs don't affect one other.

Exercise: Consider boxes containing 25 jars of baby food (from previous).

1. What is the expected weight of the boxes?
2. What is the standard deviation of the box weights?
3. Give the probability that the box weighs less than 2,975 grams.

$X \sim N(\mu = 120, \sigma^2 = 16)$, $\sigma = 4$

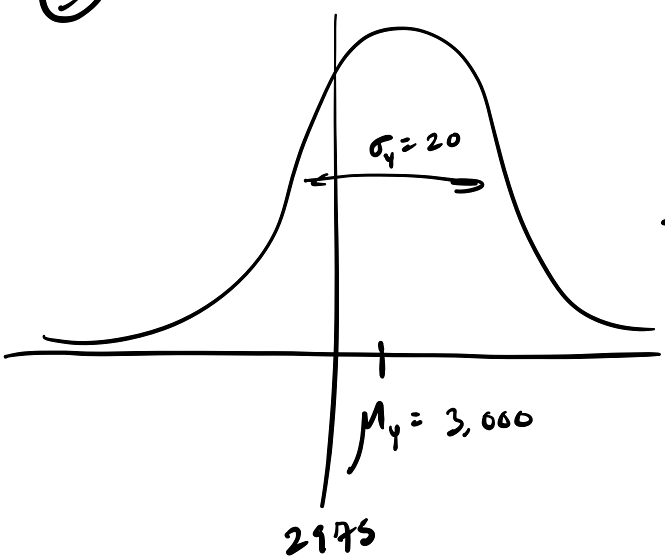
Y = weight of a box of 25 jars.

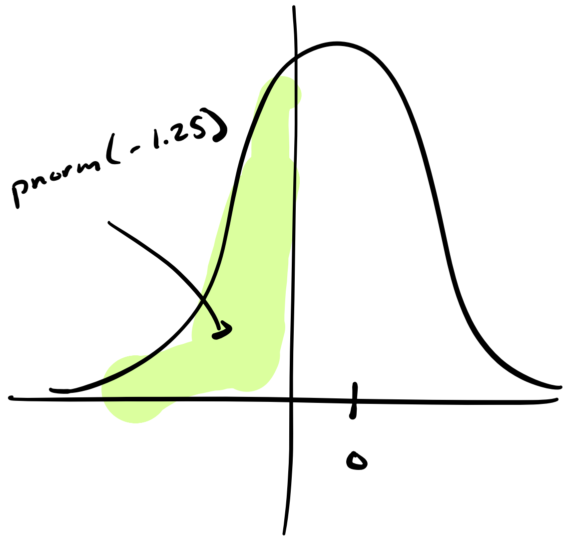
$$Y = X_1 + X_2 + \cdots + X_{25}$$

① $\mathbb{E}\,Y = 25 \cdot 120 = \mu_Y = 2500 + 25 \cdot 20 = 3{,}000$

② $\operatorname{Var} Y = 25 \cdot 16$. $\quad \sigma_Y = \sqrt{25 \cdot 16} = 5 \cdot 4 = 20.$

③





$$\frac{2975 - 3000}{20} = \frac{-25}{20} = -1.25$$