Lecture 3

Questions that we should be able to answer by the end of this lecture:

Which is the better exam score?

SD = st. dev'n

- 67 on an exam with mean 50 and SD 10

  or

- 62 on an exam with mean 40 and SD 12

Is it fair to say:

- 67 is better because 67 > 62?

  or

- 62 is better because it is 22 marks above the mean and 67 is only 17 marks above the mean?

To answer these questions we have to introduce **z-scores**.

Linear Transformations of Data

A **linear transformation** changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{new} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

Example:

(a) If a distance x is measured in kilometers, the same distance in miles is

$x_{new} = 0.6x$

10 km = 6 miles

0 km = 0 miles ← origin hasn't changed

(b) We want a temperature x measured in degrees Fahrenheit to be expressed in degrees Celsius. The transformation is

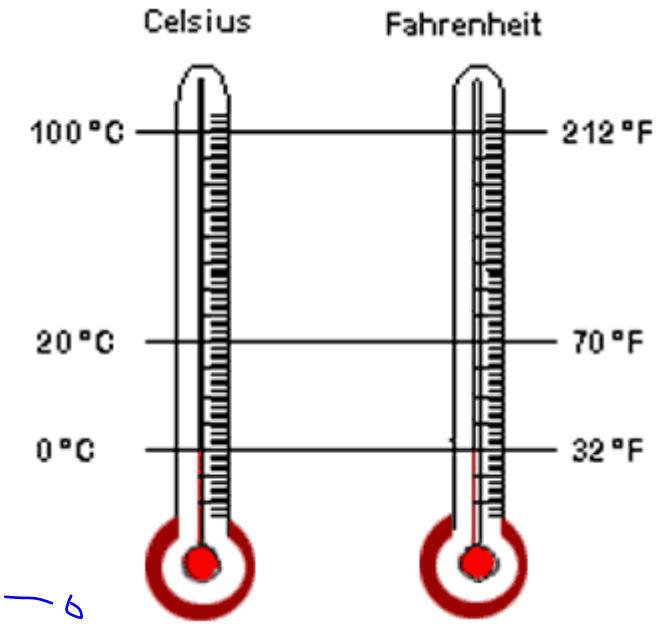
$$X_{\text{new}} = \frac{5}{9}(X - 32)$$

$$^\circ C \qquad ^\circ F$$

$$\frac{5}{9}X - 32 \cdot \frac{5}{9}$$

$$= -\frac{160}{9} + \frac{5}{9}X$$

$$a + bx$$



$$95^\circ F = -\frac{160}{9} + \frac{5}{9} = 35^\circ C$$

$$0^\circ F = -\frac{160}{9} + \frac{5}{9} \cdot 0 = -\frac{160}{9} = -17.8^\circ C$$

Note: Linear transformations do not change the shape of a distribution. origin has changed

symmetric, unimodal $\xrightarrow{\text{lin. trans.}}$ symmetric, unimodal

Effect of a linear transformation:

- Multiplying each observation by a positive number *b* multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by *b*.

- Adding the same number *a* (either positive or negative) to each observation adds *a* to measures of center and to quartiles and other percentiles but does not change measures of spread.

$X_{new} = a + bX$

$\bar{x}_{new} = a + b\bar{x}$

$M_{new} = a + bM$

$Q_{1\ new} = a + bQ_1$

$Q_{3\ new} = a + bQ_3$

$SD_{new} = b\ SD$

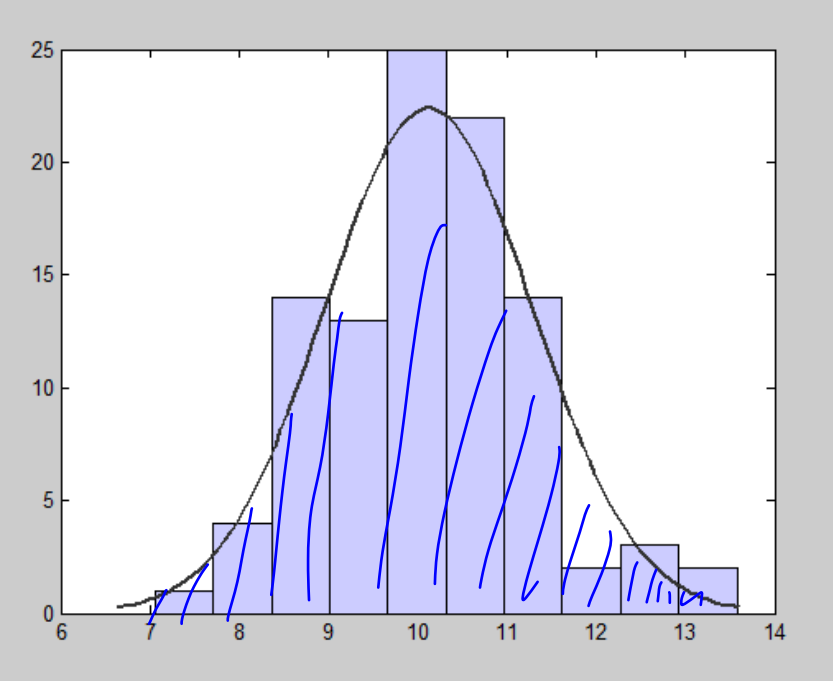
$range_{new} = b\ range$

$IQR_{new} = b\ IQR$

Note (using non-linear transformations): For skewed distributions re-express data using $\sqrt{x}$, x^2, $\ln(x)$, $\frac{1}{x}$ to get more symmetric distributions.

Density Curves

A **density curve** is a smooth approximation to the irregular bars of a histogram.



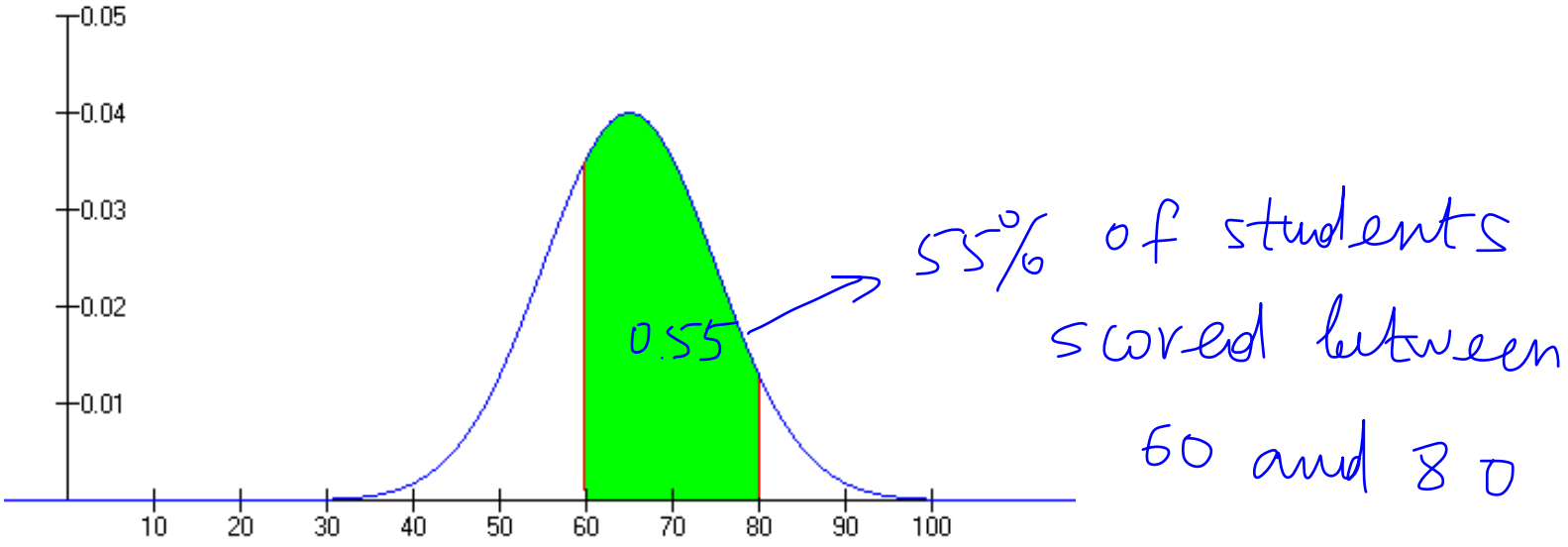
Density curve is a curve that

- is always on or above the horizontal axis.
- has area exactly 1 underneath it.

Also,

- A density curve describes the overall pattern of a distribution.
- The area under the curve and above any range of values is the proportion of all observations that fall in that range.

Example: The curve below shows the density curve for scores in an exam and the area of the shaded region is the proportion of students who scored between 60 and 80.



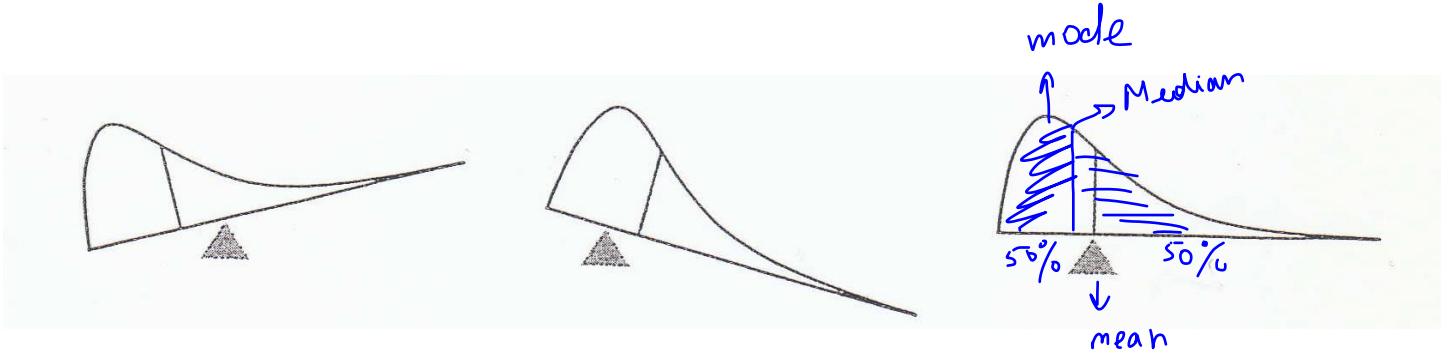
Note: Outliers are not described by the density curve.

Center and Spread for Density Curve

A **mode** of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest.

The **median** is the point with half the total area on each side.

The **mean** is the point at which the curve would balance if it were made out of solid material.



Note: The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

One can say that a density curve is an *idealized* description of a distribution of data.

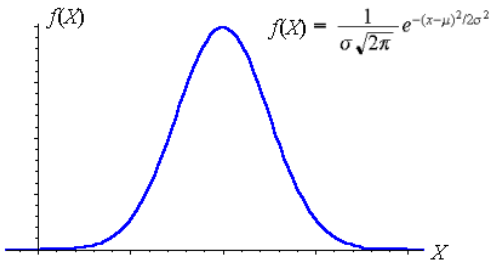
We therefore need to distinguish between the mean and standard deviation of the density curve and the numbers $\bar{x}$ and s computed from the actual observations.

We shall use the following notations:

μ (mu) – mean of density curve

σ (sigma) – st. dev. of density curve

Normal Distribution

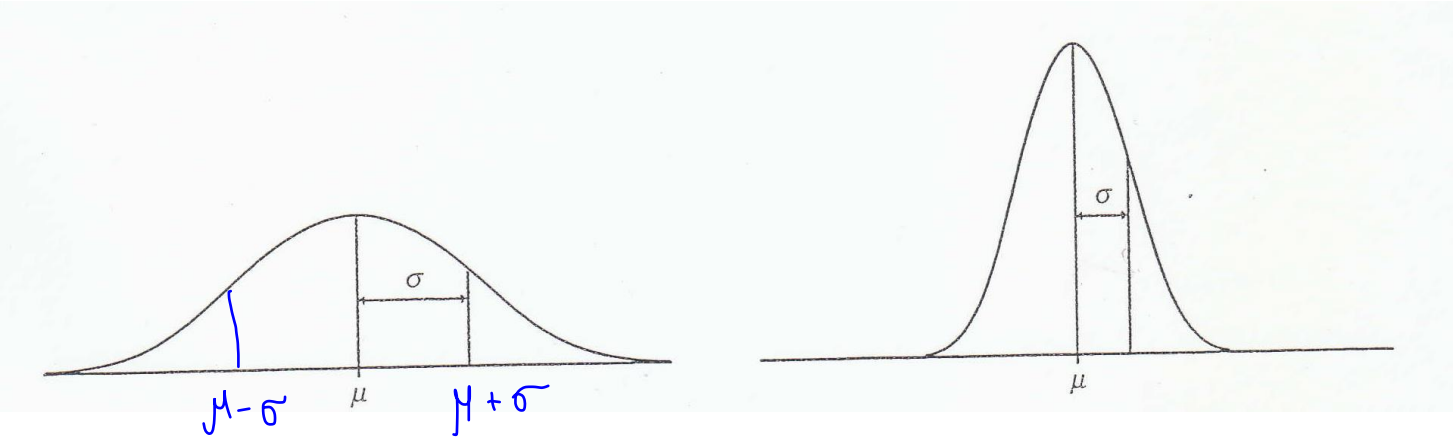


Important class of density curves: symmetric unimodal bell-shaped curves known as *normal* curves. They describe **normal distributions**:

- All normal distributions have the same overall shape.
- The density curve for a particular normal distribution is specified by giving the mean μ and the standard deviation σ.
- The mean is located at the center of the symmetric curve and is the same as the median (and mode).
- The standard deviation σ controls the spread of a normal curve.
- There are other symmetric bell- shaped density curves that are not normal.
- The normal density curves are specified by a particular equation. The height of a normal density curve at any point x is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The points at which the curve changes concavity are located at distance σ on either side of the mean μ.



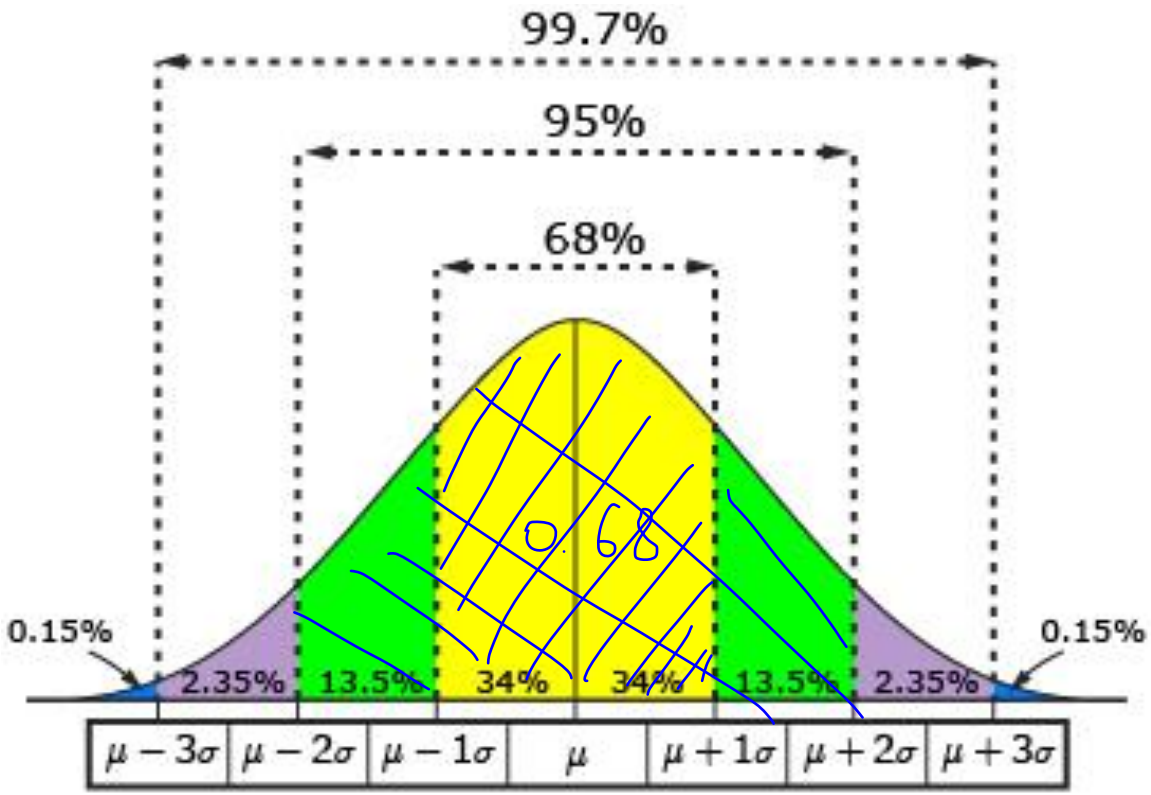
Why are the Normal distributions important?

- Normal distributions are good description for some distributions of real data.
- Normal distributions are good approximations to the results of many kinds of chance outcomes.
- Many statistical inference procedures based on Normal distributions work well for other roughly symmetric distributions.

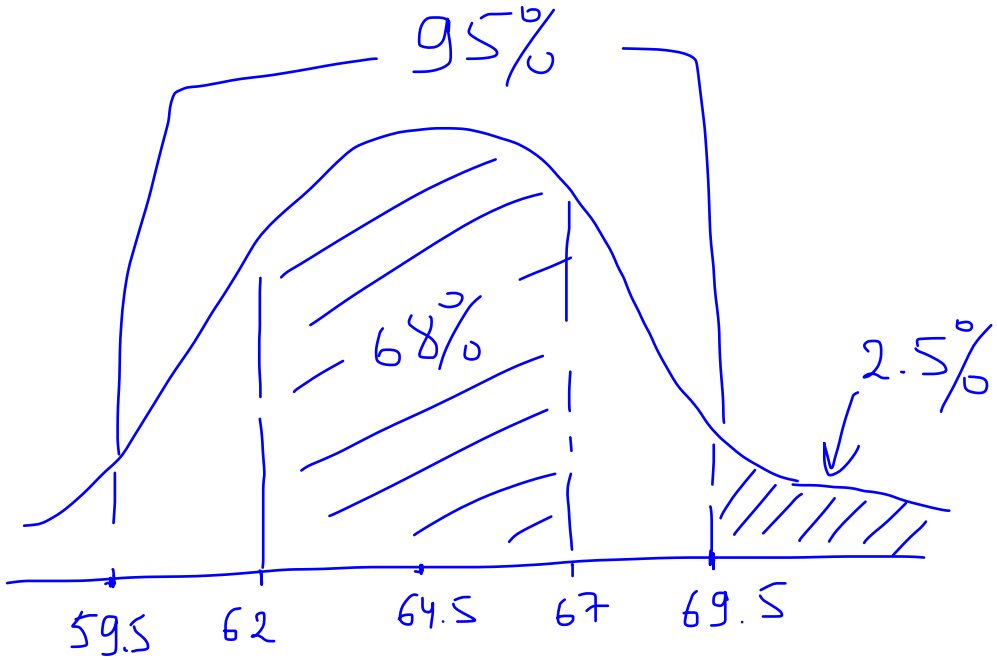
The 68-95-99.7 Rule (The Empirical Rule)

In the Normal distribution with mean μ and standard deviation σ:

- Appr.68% of the observations fall within σ of the mean μ.
- Appr. 95% of the observations fall within 2σ of the mean μ.
- Appr. 99.7% of the observations fall within 3σ of the mean μ.



Example: The distribution of heights of young women aged 18 to 24 is approximately Normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches.



$\mu \pm \sigma$

$= 64.5 \pm 2.5$

$= (62, 67)$

$\mu \pm 2\sigma = 64.5 \pm 2 \cdot 2.5 = (59.5, 69.5)$

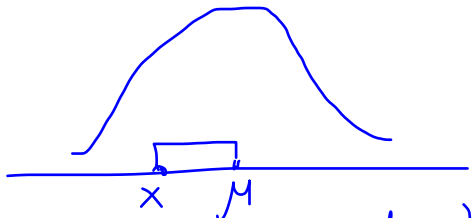
95% of women will have heights between 59.5 and 69.5

2.5% of women have heights > 69.5

Z-scores

Definition: If x is an observation from a distribution that has mean μ and standard deviation σ, the **standardized value** of x is

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



A standardized value is often called a **z-score**.

↳ tells us how many st. dev's x is from the mean

Example (heights of women, Normal with $\mu = 64.5$, $\sigma = 2.5$):

X = heights, $X \sim N(64.5, 2.5)$, μ, σ

$$z = \frac{x - 64.5}{2.5}$$

Let $x_1 = 68 \Rightarrow z = \frac{68 - 64.5}{2.5} = 1.4$

So x_1 is 1.4 st. dev's above the mean

Let $x_2 = 60 \Rightarrow z = \frac{60 - 64.5}{2.5} = -1.8$

So x_2 is 1.8 st. dev. below the mean

Back to the very first question:

Which is the better exam score?

- 67 on an exam with mean 50 and SD 10

or

- 62 on an exam with mean 40 and SD 12

Turn them into z-scores:

- 67 becomes $\frac{67-50}{10} = 1.7$

- 62 becomes $\frac{62-40}{12} = 1.83 > 1.7$

Conclusion: 62 is (slightly) better relative to the mean and SD

Definition: The **standard Normal distribution** is the Normal distribution $N(0,1)$ with mean 0 and standard deviation 1.

Note: If a variable X has any Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the standardized variable

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

has the standard Normal distribution. Why?

$$Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$$

$$a + bX$$

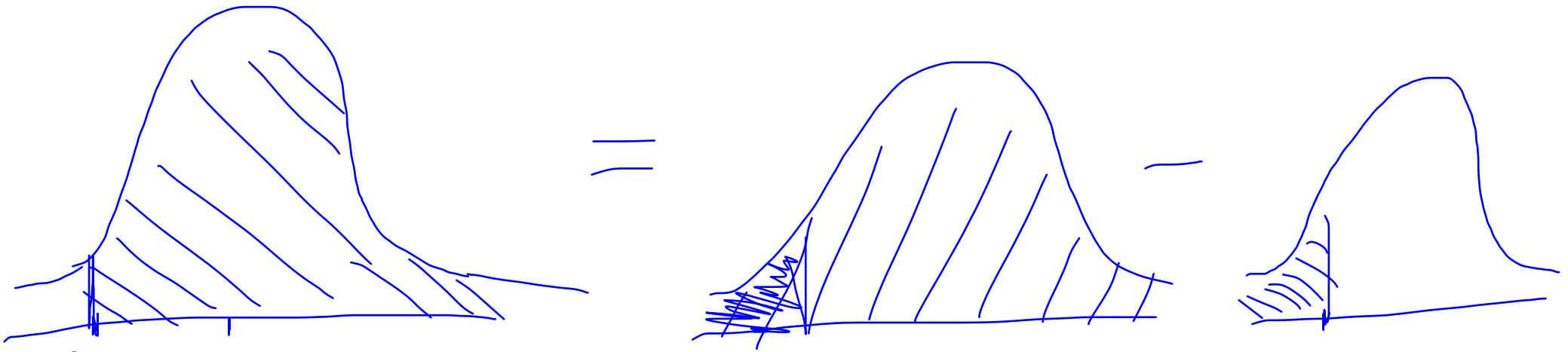
$$\mu_Z = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\cdot\mu = 0 \leftarrow \text{mean of new variable } Z$$

$$\sigma_Z = \frac{1}{\sigma}\cdot\sigma = 1 \rightarrow \text{st. dev. of new variable } Z$$

Example: The National Collegiate Athletic Association (NCAA) requires Division I athletes to get a combined score of at least 820 on SAT Mathematics and Verbal tests to compete in their first college year. (Higher scores are required for students with poor high school grades.) The scores of the 1.4 million students in the class of 2003 who took the SATs were approximately Normal with mean 1026 and standard deviation 209. What proportion of all students had SAT scores of at least 820?

Let $X = \text{scores} \sim N(1026, 209)$

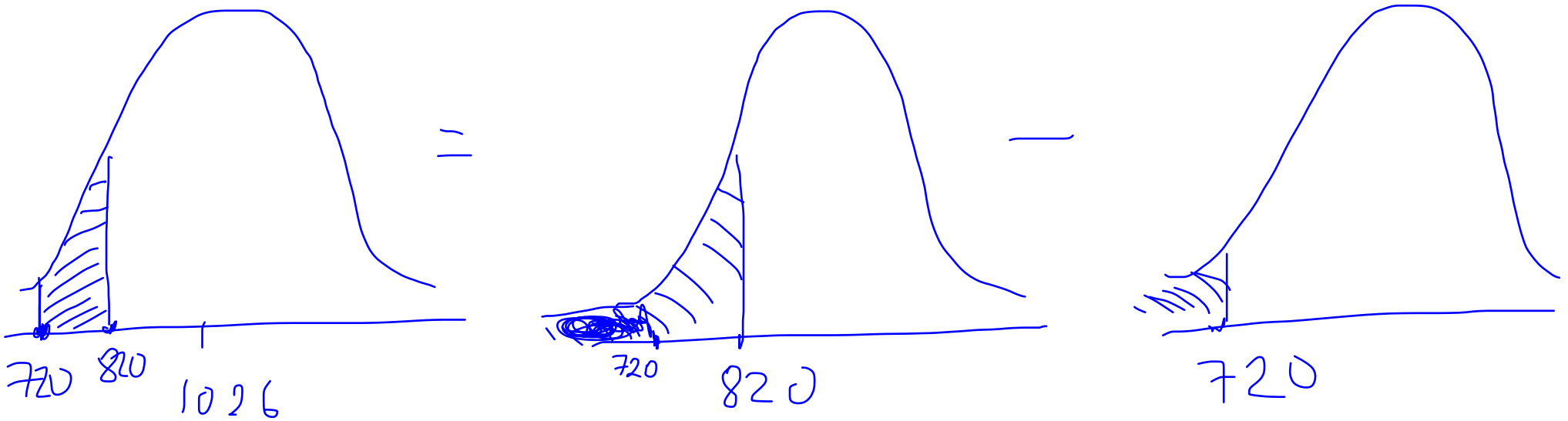
$\text{Prop}(X \geq 820) = P(X \geq 820)$



$P(X \geq 820) = 1 - P(X \leq 820)$

The NCAA considers a student a «partial qualifier» eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. What proportion of all students who take the SAT would be partial qualifiers? That is, what proportion have scores between 720 and 820?

$$P(720 \le X \le 820)$$



$$P(720 \le X \le 820) = P(X \le 820) - P(X < 720)$$

Standard Normal Table (Z-Table)

To find the proportions for the above example:

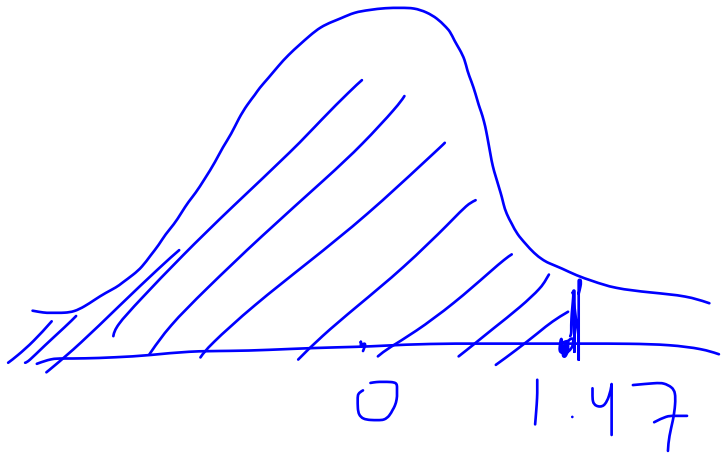
- Standardize
- Use Z-Table

Example: What proportion of observations of a standard Normal variable Z take values less than 1.47?

$Z \sim N(0,1)$

$P(Z < 1.47)$

$= 0.9292$ ← from Z-table



Example: What proportion of all students who take the SAT have scores of at least 820?

$$P(X \geq 820) = 1 - P(X < 820)$$

$X \sim N(1026, 209)$

Z-score for 820 $= \frac{820 - 1026}{209} = -0.99$

So $P(X \geq 820) = 1 - P(X < 820)$

$= 1 - P(Z < -0.99)$

$= 1 - 0.1611 = 0.3389$

$\approx 84\%$ of students scored at least 820

Example: What proportion of students have SAT scores between 720 and 820?

$$P(720 \le X \le 820) = P(X \le 820) - P(X < 720)$$

$$\text{z-score for } 820 = \frac{820 - 1026}{209} = -0.99$$

$$\text{z-score for } 720 = \frac{720 - 1026}{209} = -1.46$$

$$= P(Z \le -0.99) - P(Z < -1.46)$$

$$= 0.1611 - 0.0721 = 0.0890$$

Inverse Normal Calculations

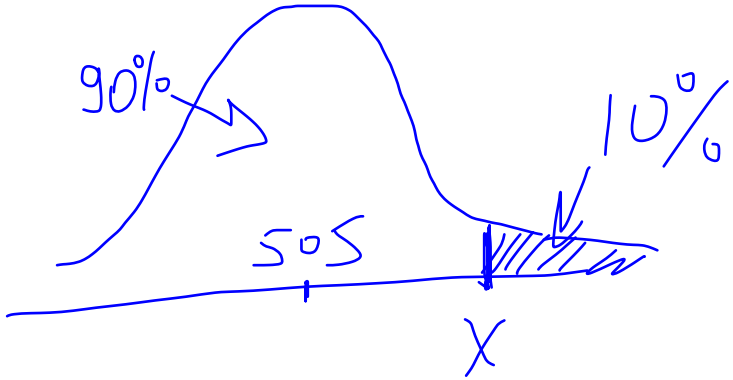
- Use Z-Table backwards to get Z
- Turn Z back into original scale by using $X = \mu + \sigma \cdot Z$

Why? $Z = \frac{X-\mu}{\sigma} \Rightarrow \sigma Z = X - \mu \Rightarrow X = \mu + \sigma Z$

Example: Scores on the SAT Verbal test in recent years follow approximately the $N(505, 110)$ distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

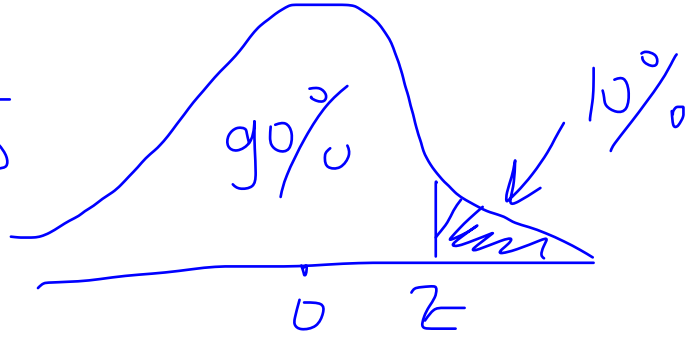
$X \sim N(505, 110)$

$Z \approx 1.285$ (average of 1.28 and 1.29)



$X = \mu + \sigma Z = 505 + 110 \cdot 1.285$

$= 646.35$



Normal Quantile Plots

A histogram or stemplot can reveal distinctly non-normal features of a distribution.
If the stemplot or histogram appears roughly symmetric and unimodal, we use another graph, the **Normal quantile plot** as a better way of judging the adequacy of a Normal model

Here is the basic idea of this plot:

- Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies.

  For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.

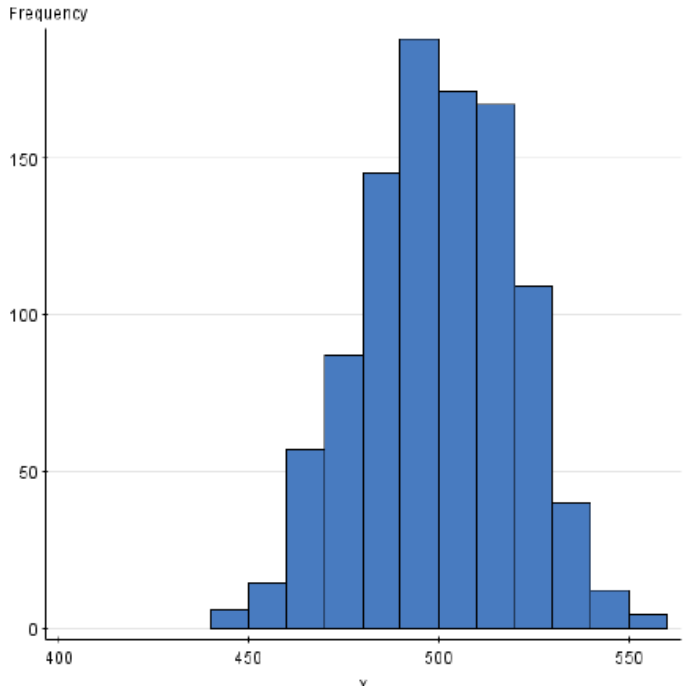
- Do Normal distribution calculations to find the values of z corresponding to these same percentiles.

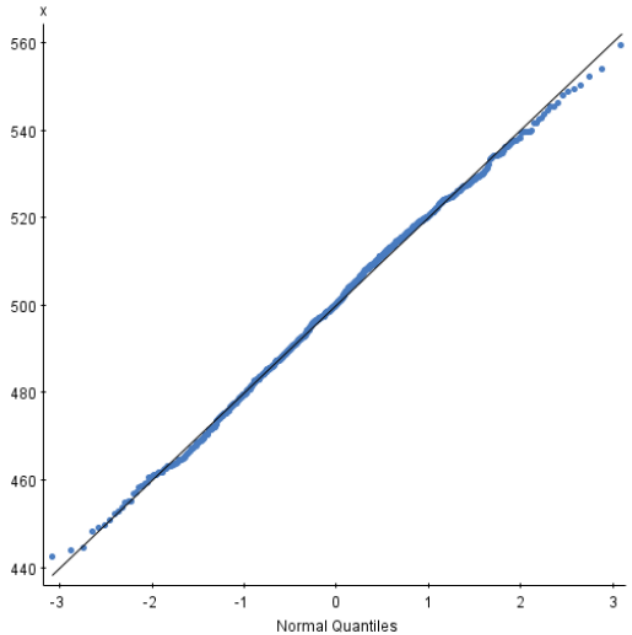
  For example, $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point. We call these values of z **Normal scores (nscores)**.

- Plot each data point *x* against the corresponding Normal scores. If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

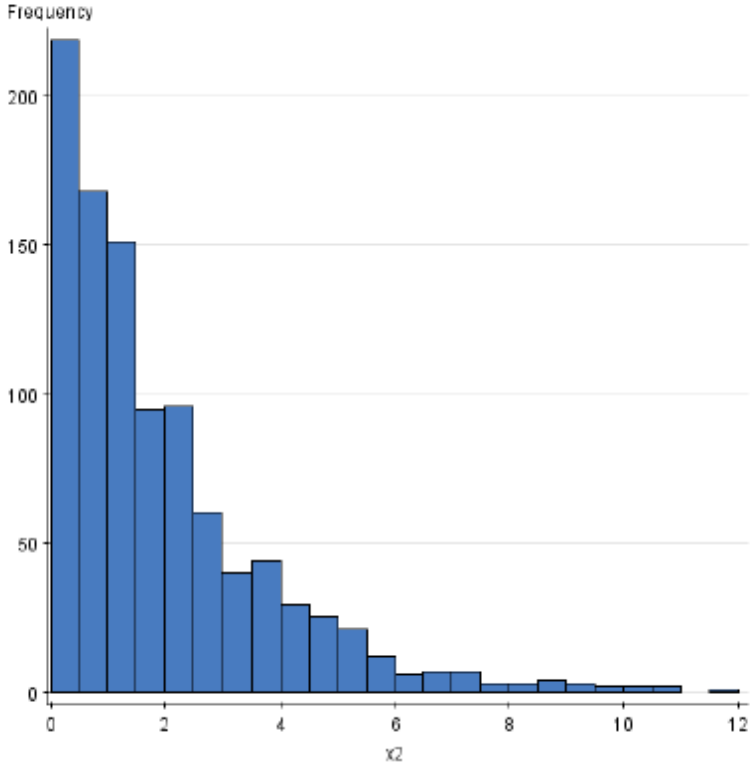
Examples:

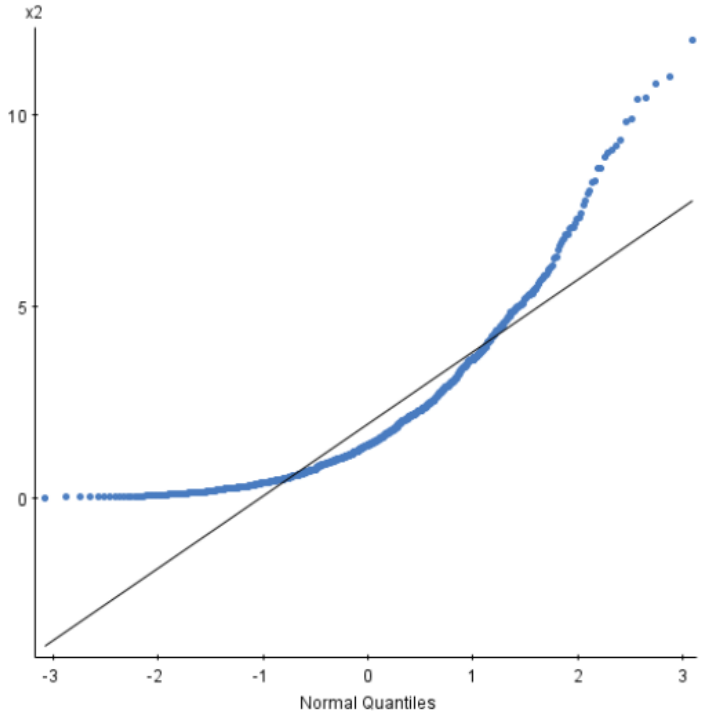
- Histogram and the nscores plot for data generated from a normal distribution (N(500, 20)) (for 1000 observations)



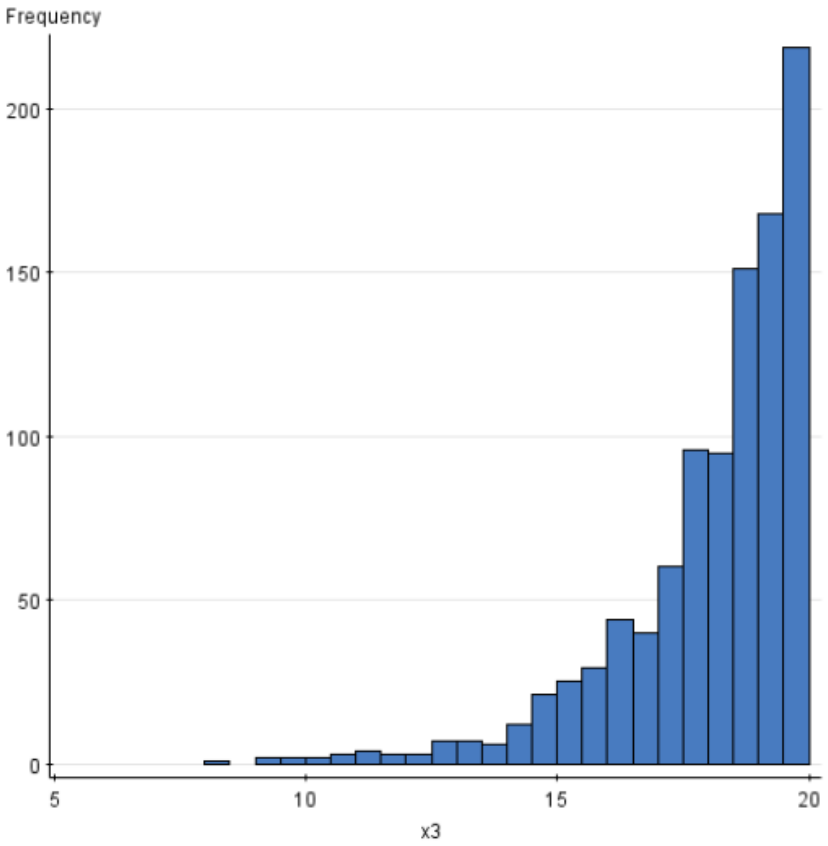


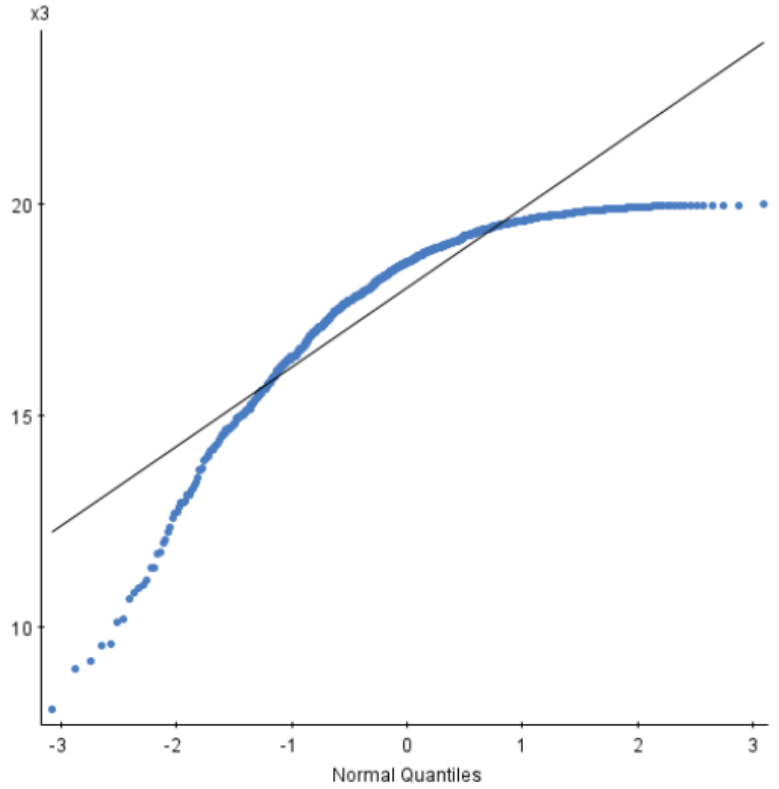
- Histogram and the nscores plot for data generated from a right skewed distribution.





- Histogram and the nscores plot for data generated from a left skewed distribution





StatCrunch -> Graphics -> QQ Plot