The Normal Distribution!

STAT 120

Day 14

No new concepts!

- We've covered the core ideas of intro stats:
  - EDA: using pictures and numbers to make sense of data
  - Estimation: estimating unknown parameters with confidence
  - Testing: assessing hypotheses with p-values
- Rest of the course covers more types of inference methods
  - Instead of using computer simulations to generate bootstrap/randomization distributions
  - Use probability models to do this

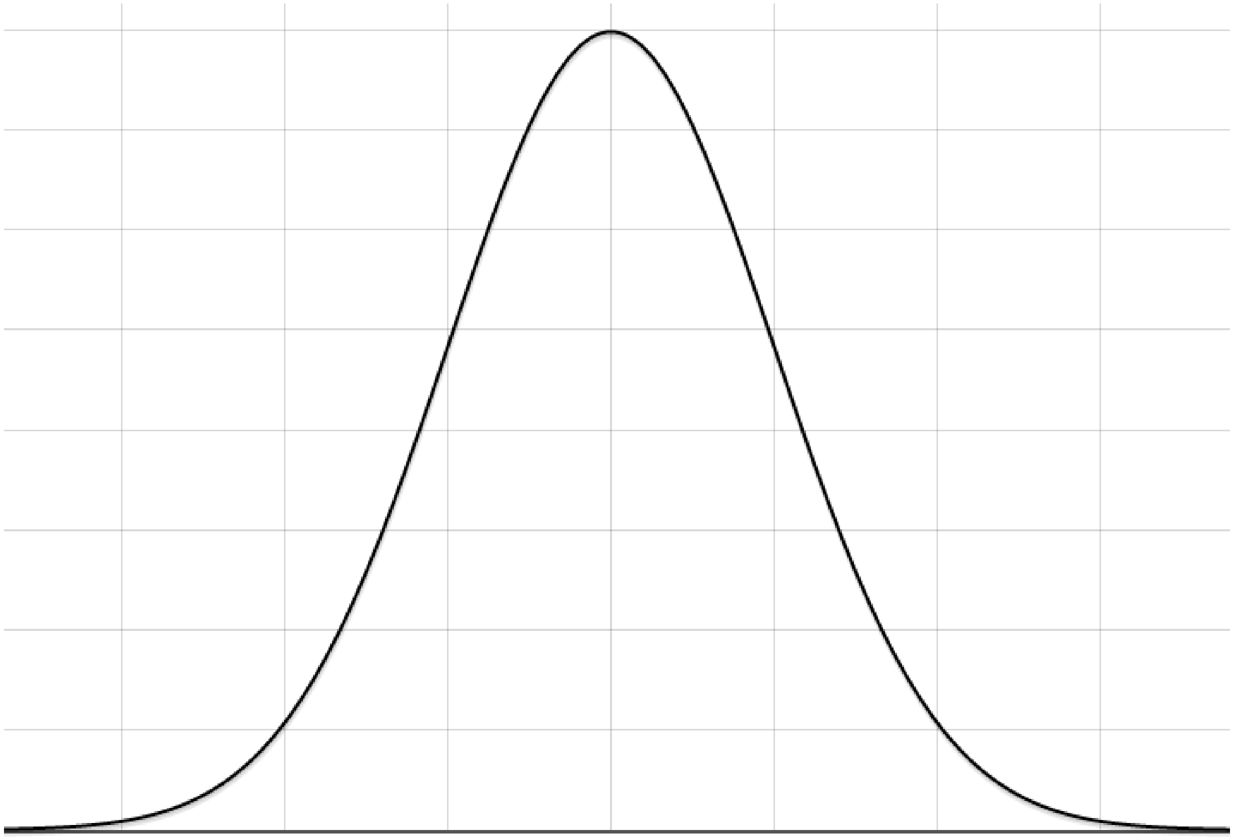
Density Curve

A ***density curve*** is a theoretical model to describe a distribution.

- Distribution for
  - **Individual** measurements in **population** (for a quantitative variable)
  - **Sampling distribution** for a **statistic**
- All density curves:
  - have an area under the curve of 1 (100%)
  - **give proportions/percents as areas under the curve**

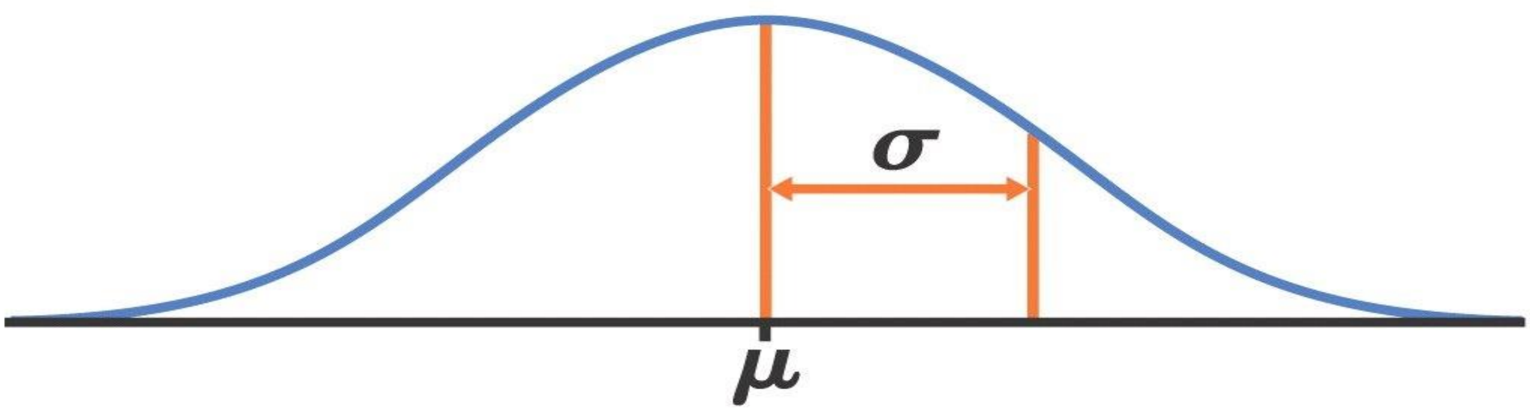
Normal Distribution

A ***normal distribution*** has a symmetric bell-shaped density curve.



The Normal Model: $X \sim N(\mu, \sigma)$

- The mean and SD determine how a normal density curve looks.
- The normal model **parameters** are
  - μ = model **mean (center)**
  - σ = model **SD (variability)**

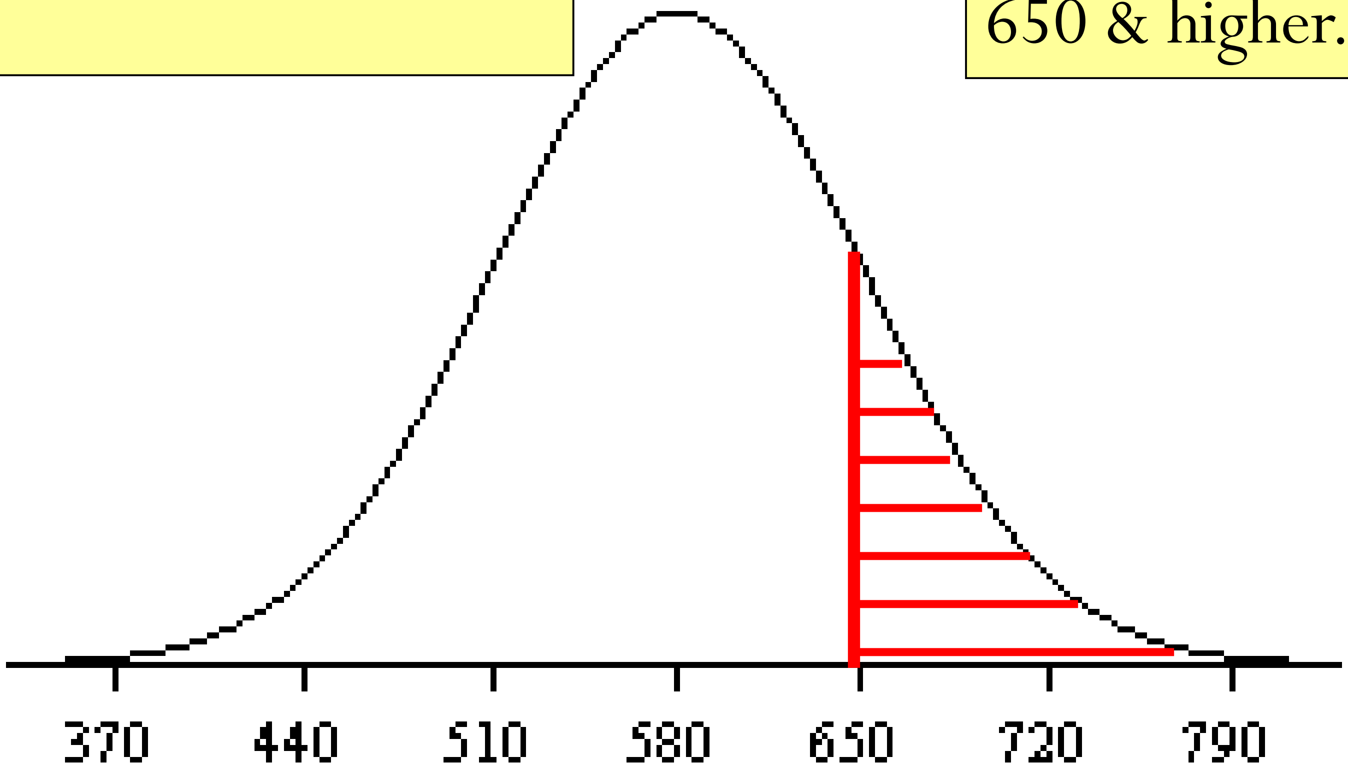


Example: A Population

Verbal SAT ~ N(580, 70)

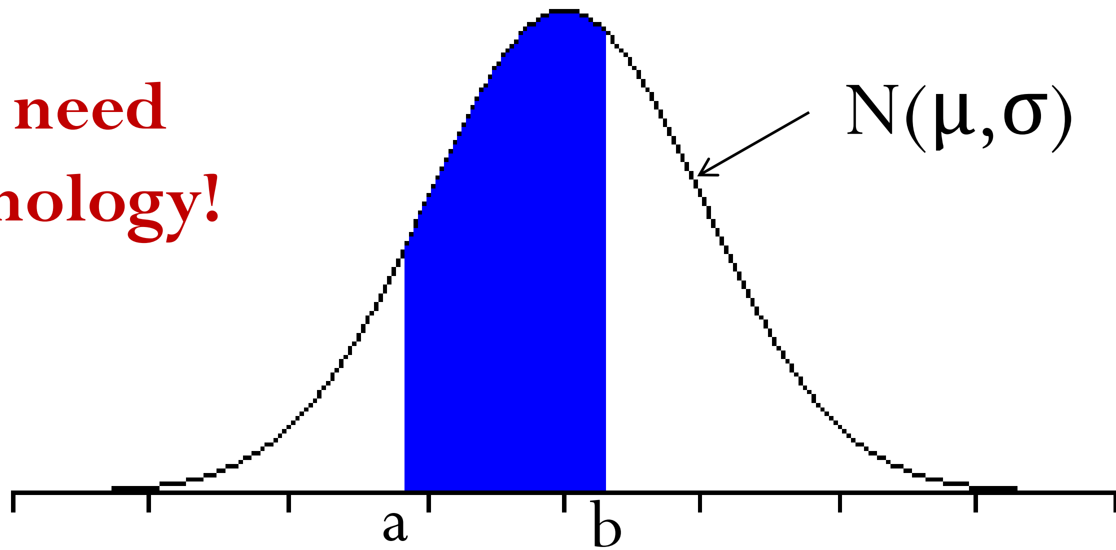
What proportion of people score above 650?

Area under the curve from scores 650 & higher.



How can we find areas under a normal density?

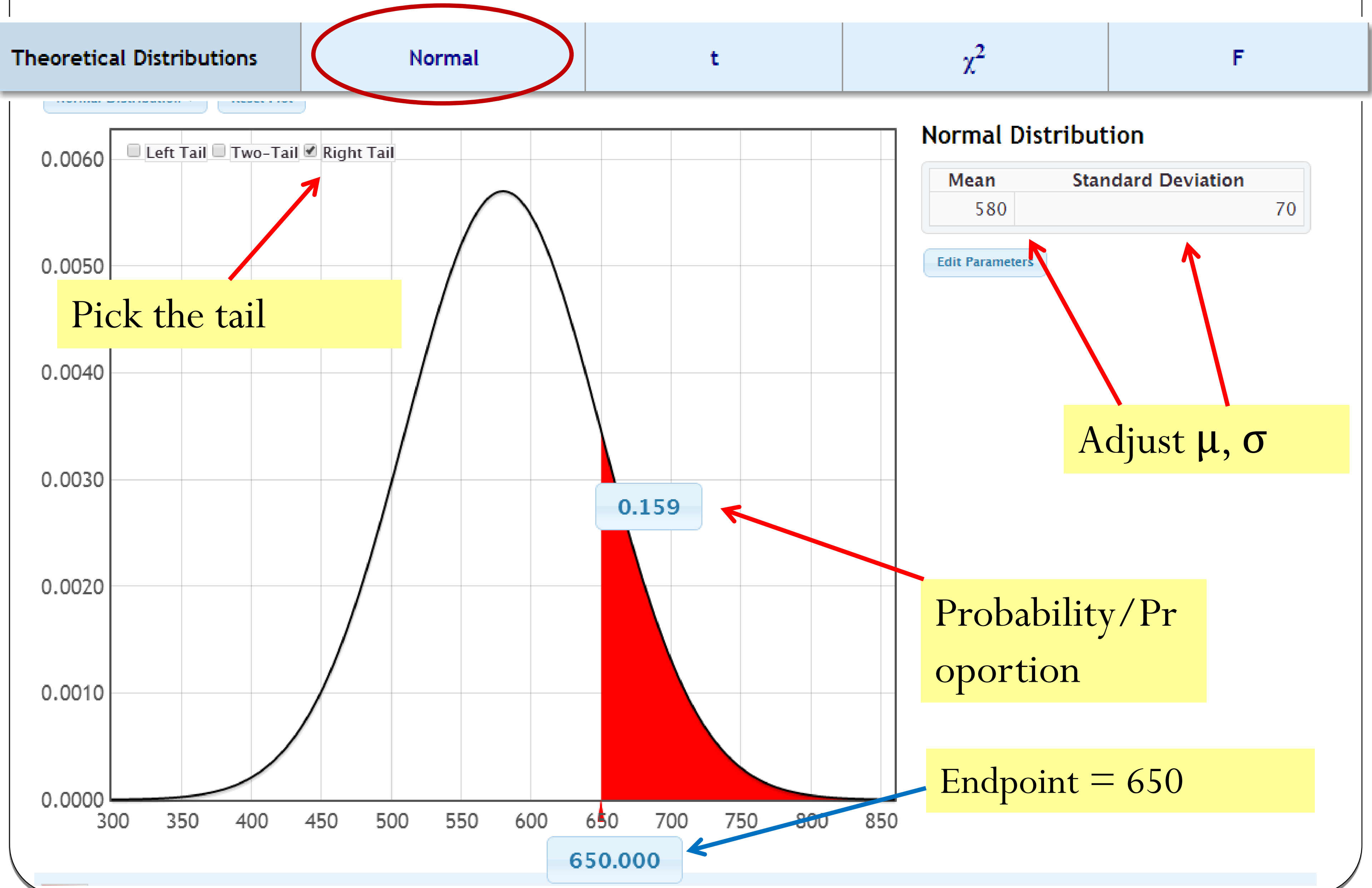
We need technology!



Calculus!

$$Area = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

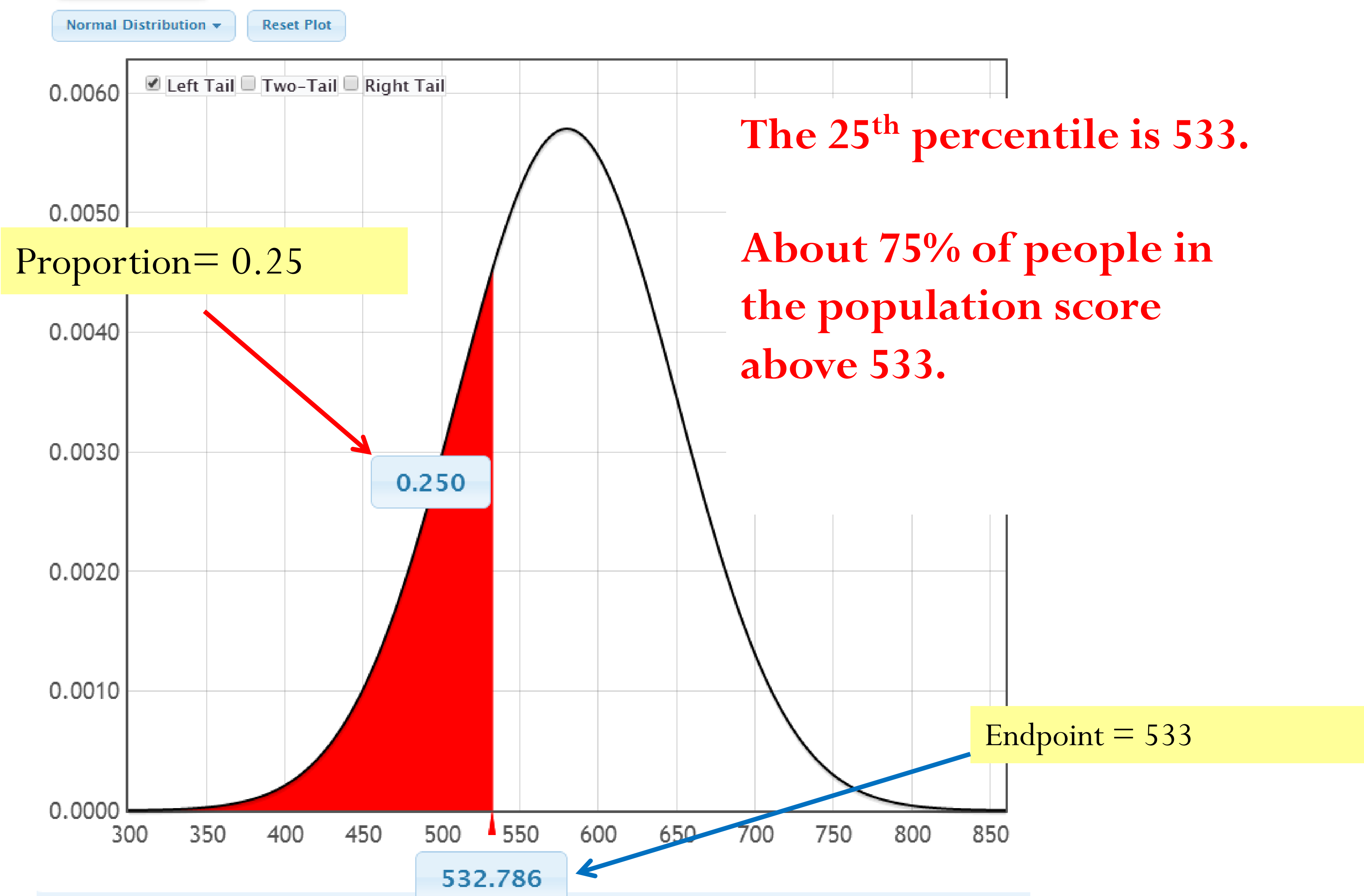
(1a) StatKey – Verbal SAT population



Example: Verbal SAT scores

- About 16% of people in this population had a score of 650 or higher.
- What score is the 25th percentile?
  - without Statkey – it will be a score below 580 (median & mean) and above 440 (2 SD below 580)
  - using Statkey – adjust the left-tail area to be 0.25.

(1b) StatKey – Verbal SAT population



The 25th percentile is 533.

About 75% of people in the population score above 533.

Example: Verbal SAT scores

- What percent of the population had a score of 650 or higher?
- Using R enter:

> 1 – pnorm(650, 580,70)

What score is the 25th percentile?

- Using R enter:

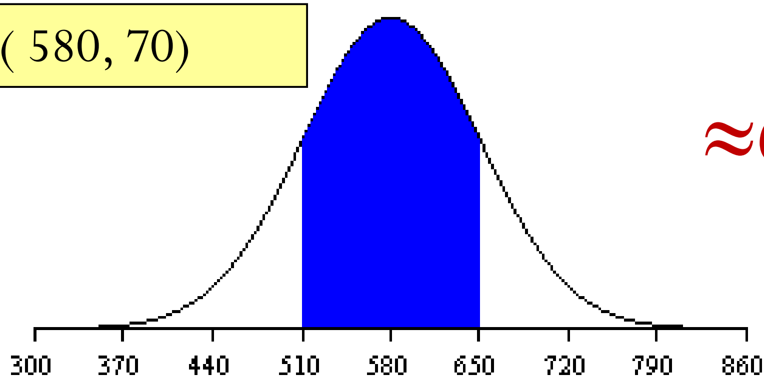
> qnorm(.25,580,70)

Finding Probabilities for N(μ,σ)

Big Idea for Normal Models: **All that really matters is the number of standard deviations from the mean.**

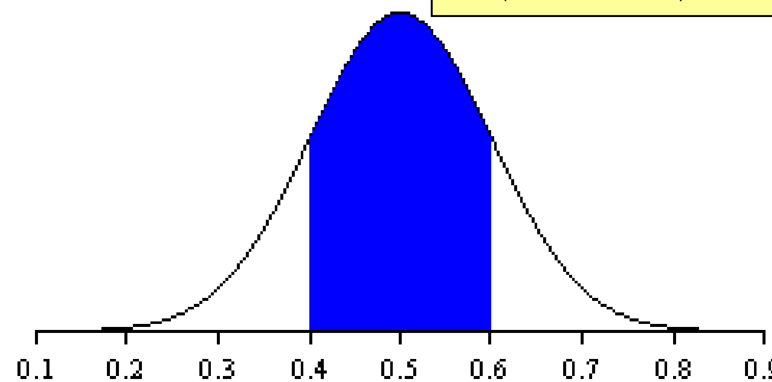
About what proportion should be within one standard deviation of the mean?

N(580, 70)

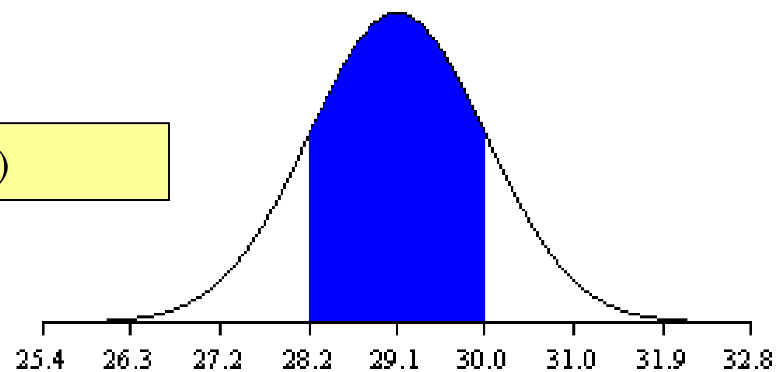


≈68%

N(0.5, 0.1)



N(29.1, 0.9)



Big Idea for Normal Models:

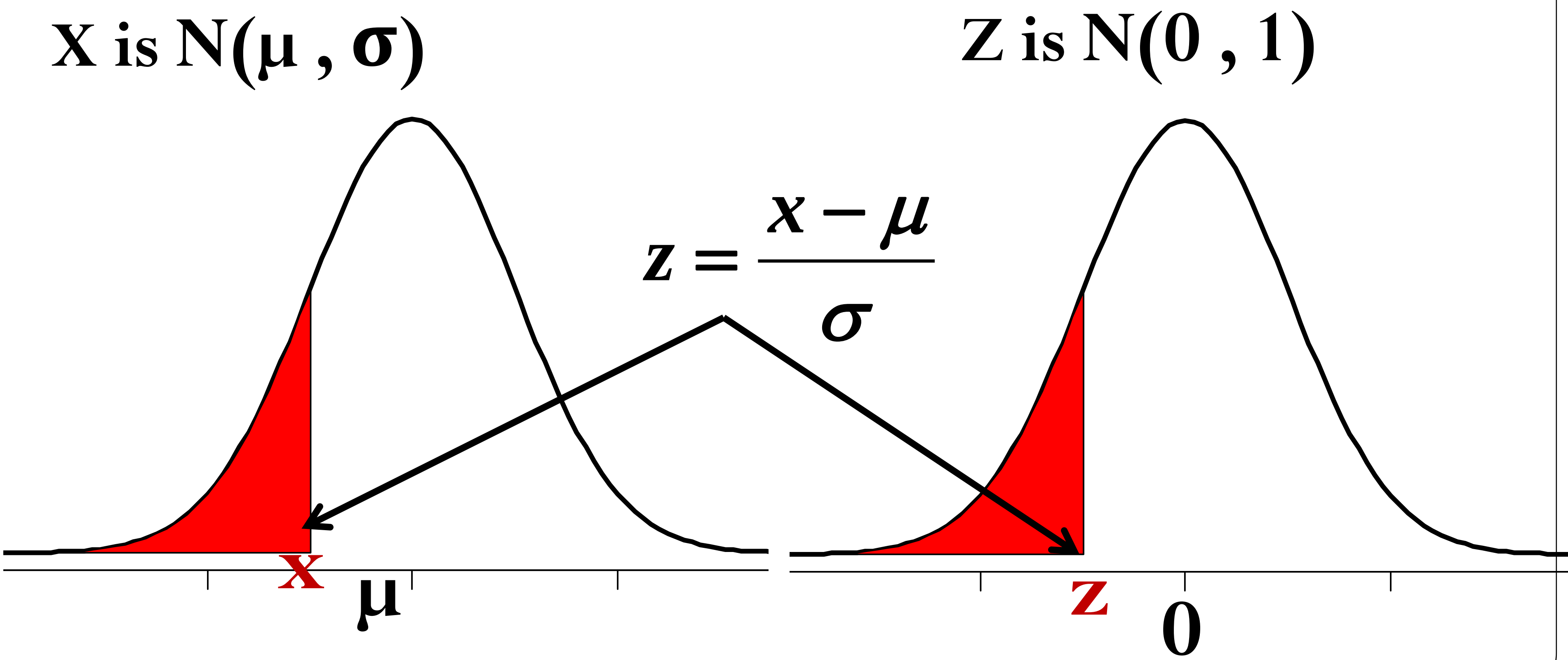
All we need is a z-score.

Standard Normal

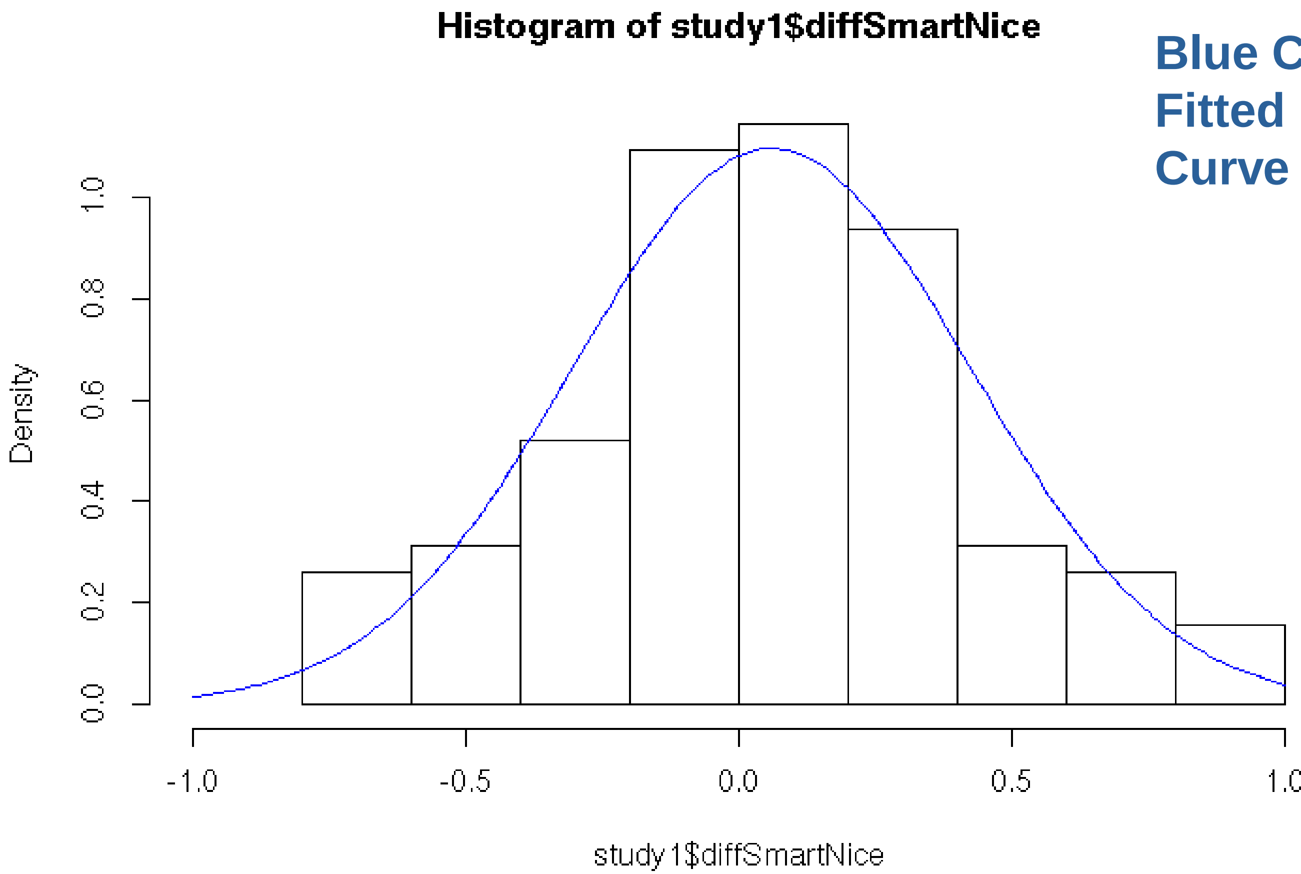
$$\boxed{\mu=0,\ \sigma=1 \rightarrow Z\sim N(0,1)}$$

Connecting any Normal model to the standard normal model

Area below x = Area below z



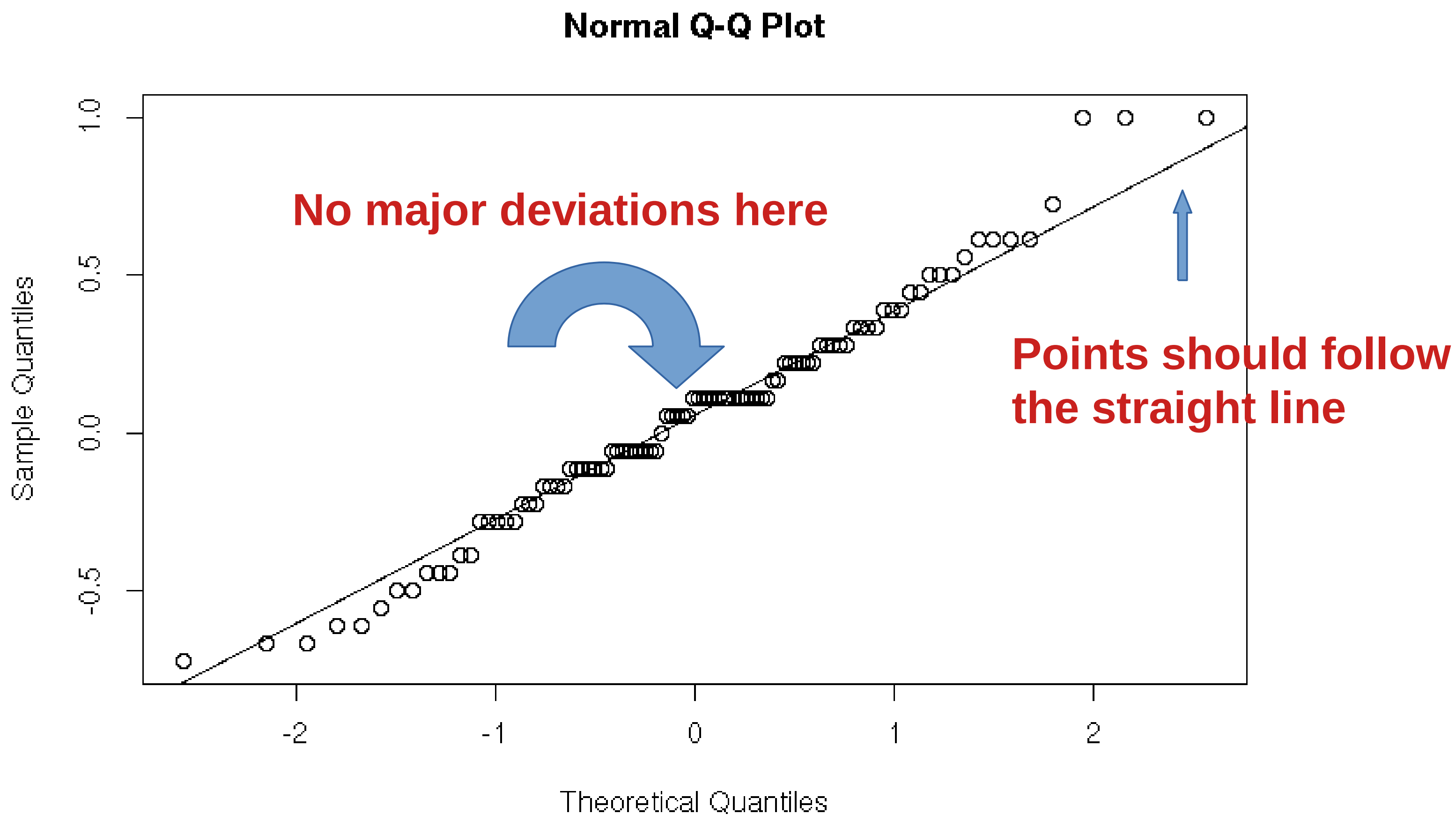
Normal Approximation: Gender Difference Study 1



Blue Curve: Fitted normal Curve

> dnorm(x-values, mean, sd)

QQ-plot: Gender Difference Study 1



The difference between nice and smart scores follows a normal distribution.

Big question

- When have we already been using normal models??
  - Bootstrap distributions – get confidence intervals if a bootstrap distribution is roughly bell-shaped
  - Randomization distributions – many of these are bell-shaped.
- Normal models play a huge role in statistical inference.
- **If we know the (bootstrap/randomization) standard error*** then we can just use a normal model rather than a resampling model (which requires more computational effort).