T-Statistics

- want to estimate a population mean μ for a normal distribution $N(\mu, \sigma)$ when we don't know μ or σ

  Example: distribution of weights in our box

- get a sample and measure

  Example: sample of 5 values from our box of weights:
  120 165 110 113 185

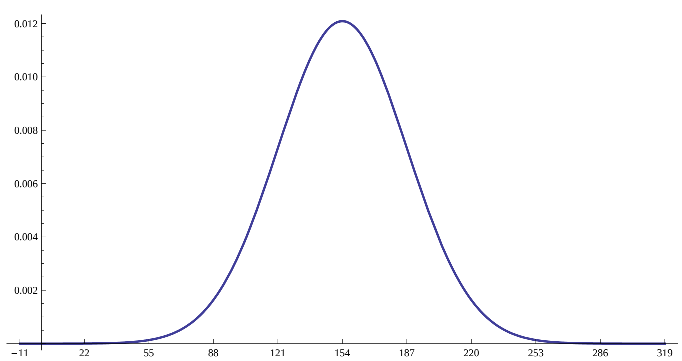
- compute mean $\bar{x}$ and standard deviation s for the sample

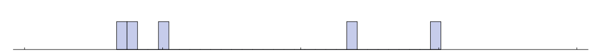
  Example: $\bar{x} = \dfrac{120 + \cdots + 185}{5} = 138.6$

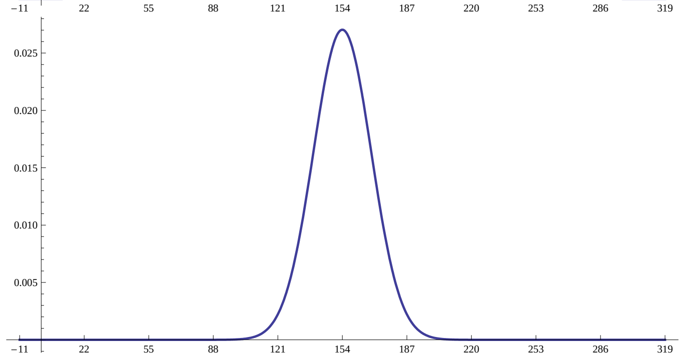
  $$s = \sqrt{\frac{(120 - 138.6)^2 + \cdots + (185 - 138.6)^2}{5 - 1}} = 34.2$$

- have three distributions in play
  - the population distribution $N(\mu, \sigma)$
  - the data distribution with mean $\bar{x}$ and standard deviation s
  - the sample means distribution $N(\mu_{\bar{x}}, \sigma_{\bar{x}}) = N(\mu, \sigma/\sqrt{n})$

  Example: For our box of weights: $\mu = 154$, $\sigma = 33$, have



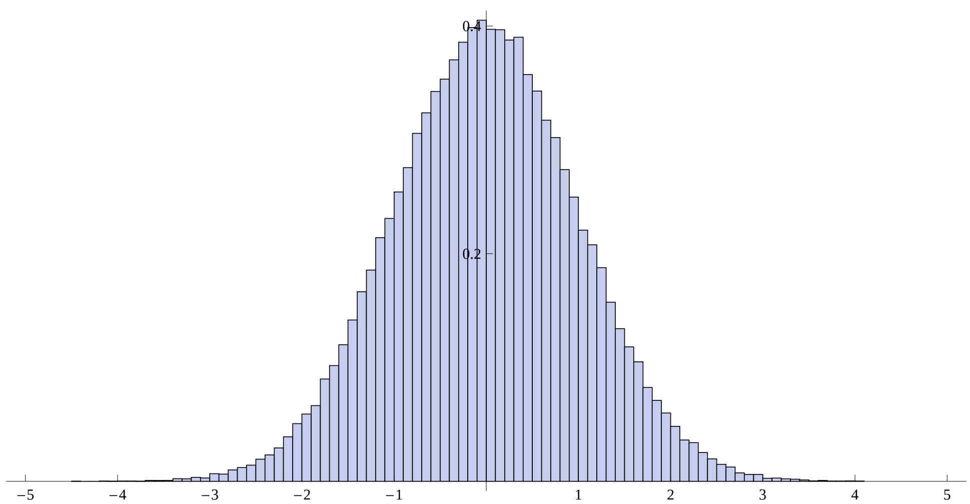


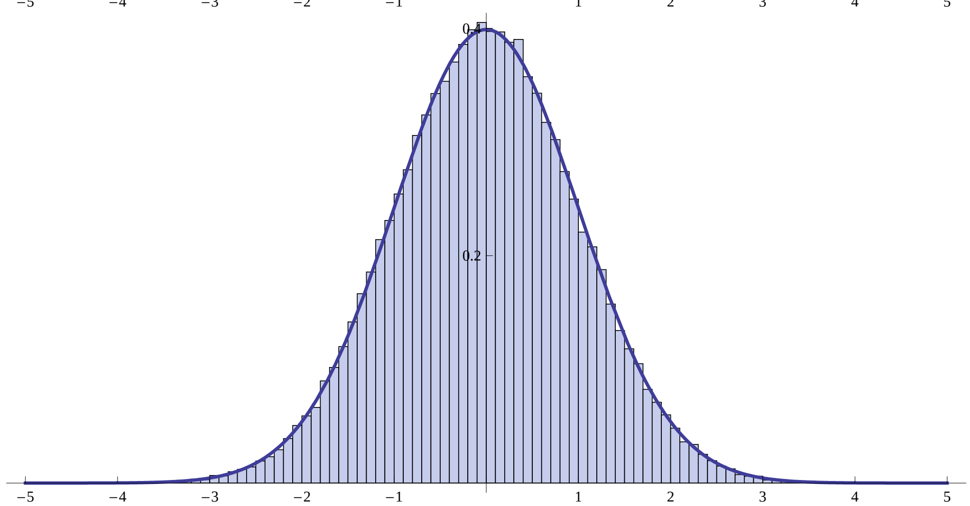


- use $\bar{x}$ as estimate of μ
- how is the estimate affected by variability from sample to sample?
- to understand, imagine we do know μ and σ
- for each possible sample mean $\bar{x}$, can compute $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
- know that z-scores have $N(0,1)$ distribution
- can check this with a sampling simulation

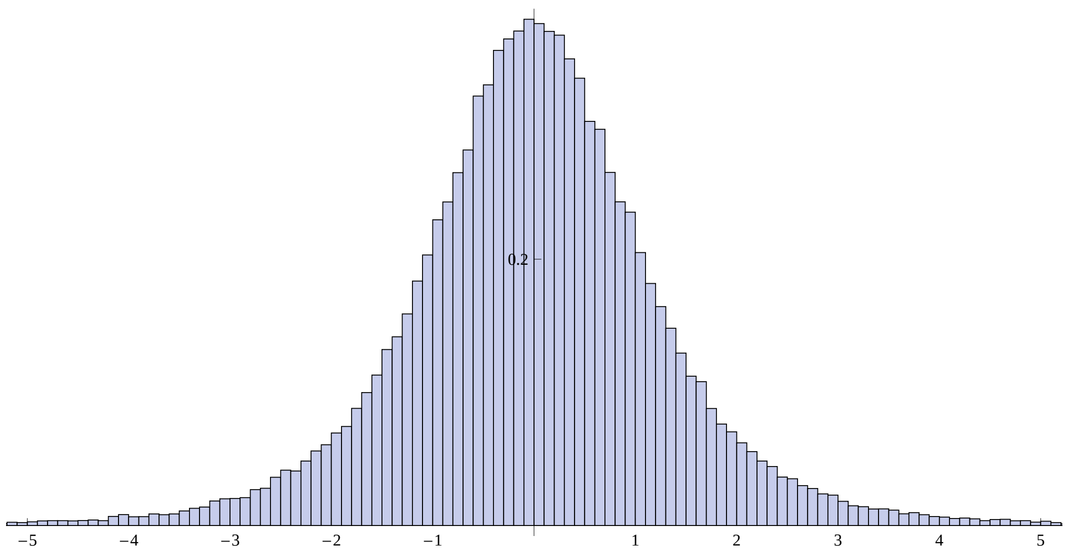
  Example: $\mu = 154$ and $\sigma = 33$

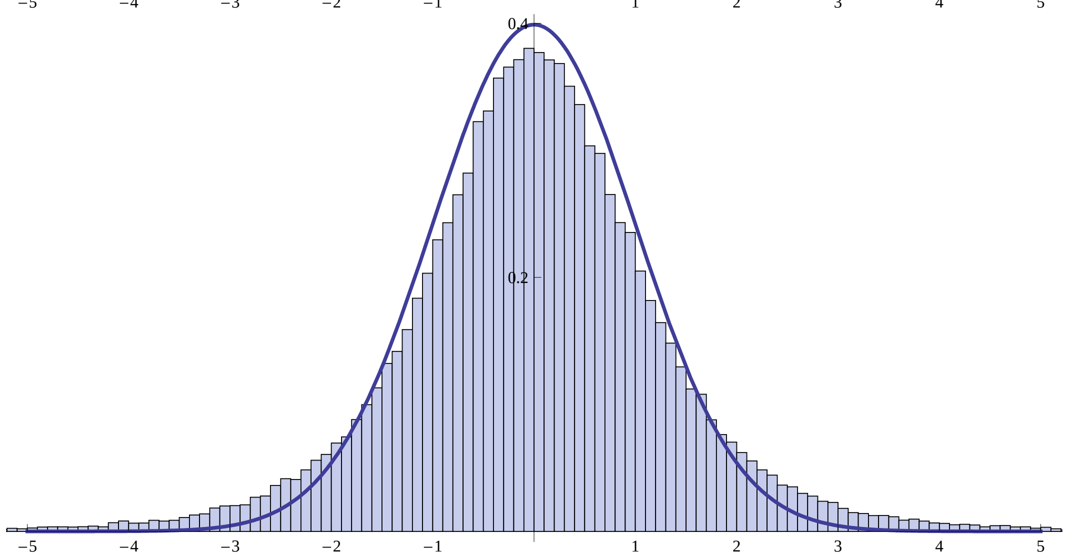
  - have software take sample of size 5 from $N(154, 33)$
  - compute $\bar{x}$ and z
  - repeat to get z for many samples of size 5
  - make histogram of these z values
  - result looks normal; confirm by superimposing plot of $N(0,1)$





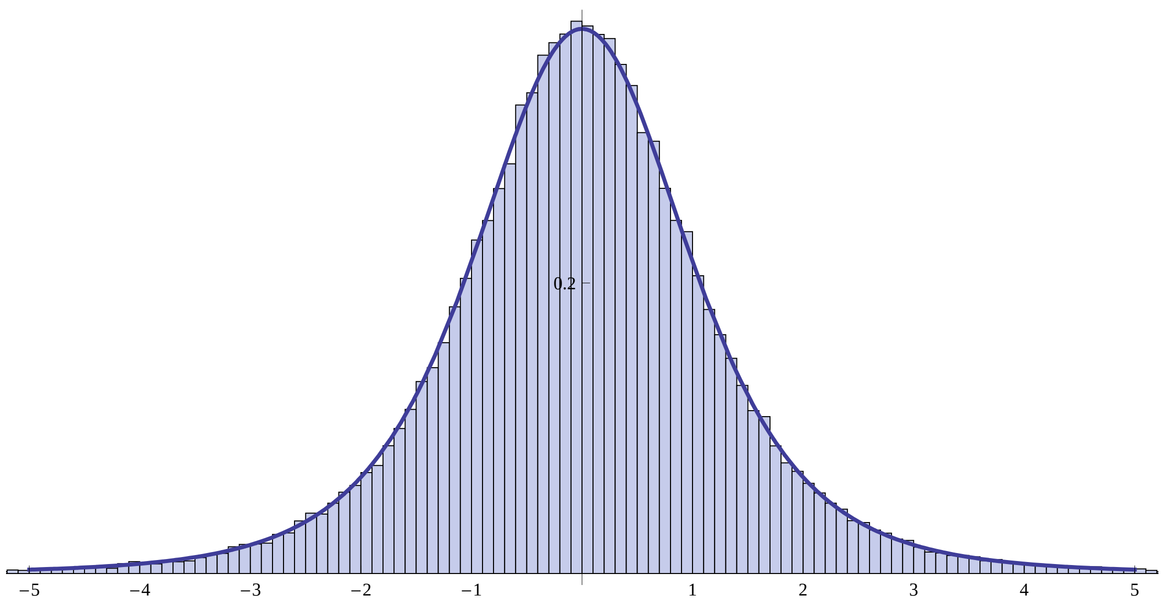
- what is the impact of using s in place of σ?
- in place of z values, we compute $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$
- to understand distribution of t values, can do a sampling simulation
  - pick values $\mu = 154$ and $\sigma = 33$
  - have software take sample of size 5 from $N(154, 33)$
  - compute $\bar{x}$, s, and t
  - repeat to get t for many samples of size 5
  - make histogram of these t values
  - result does not look normal
  - confirm by superimposing plot of $N(0, 1)$



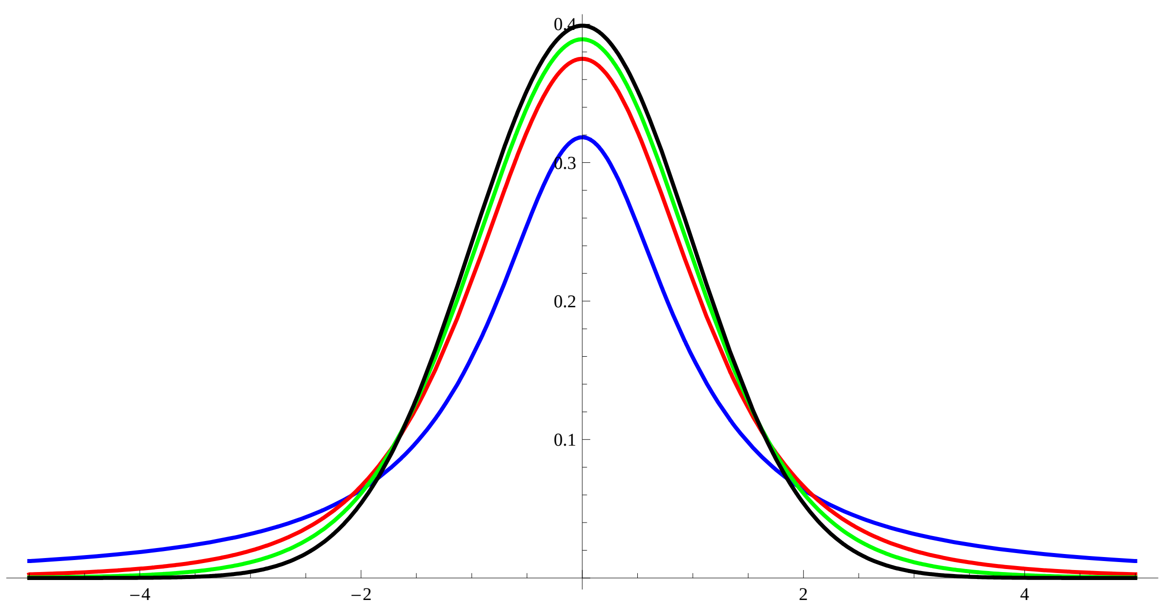


- $N(0, 1)$ is not a good fit for the distribution of t values in this simulation

- need a new type of distribution: t-distributions
- not one, but many labeled by *degrees of freedom df*
- for t values from samples of size n, use t-distribution of degrees of freedom $df = n - 1$
- for our simulation, $n = 5$, so superimpose t-distribution with $df = 5 - 1 = 4$ to compare



- t-distributions have more area in tails in comparison with $N(0, 1)$
- $t(df)$ gets closer to $N(0, 1)$ as df gets larger



Plot of $t(1)$ (blue curve), $t(4)$ (red curve), $t(10)$ green curve, $N(0, 1)$ (black curve)