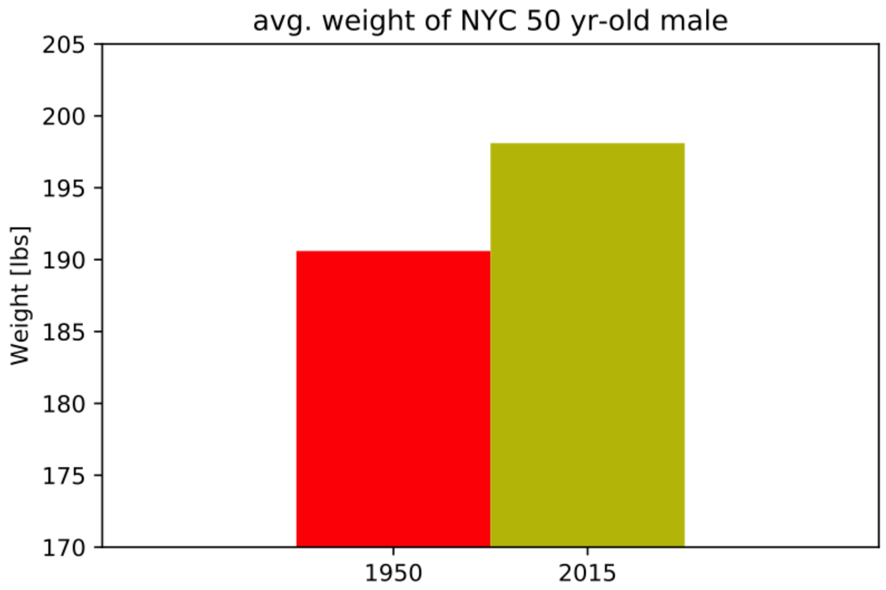
Douglas Rubin

Class 2 of Stats Bootcamp

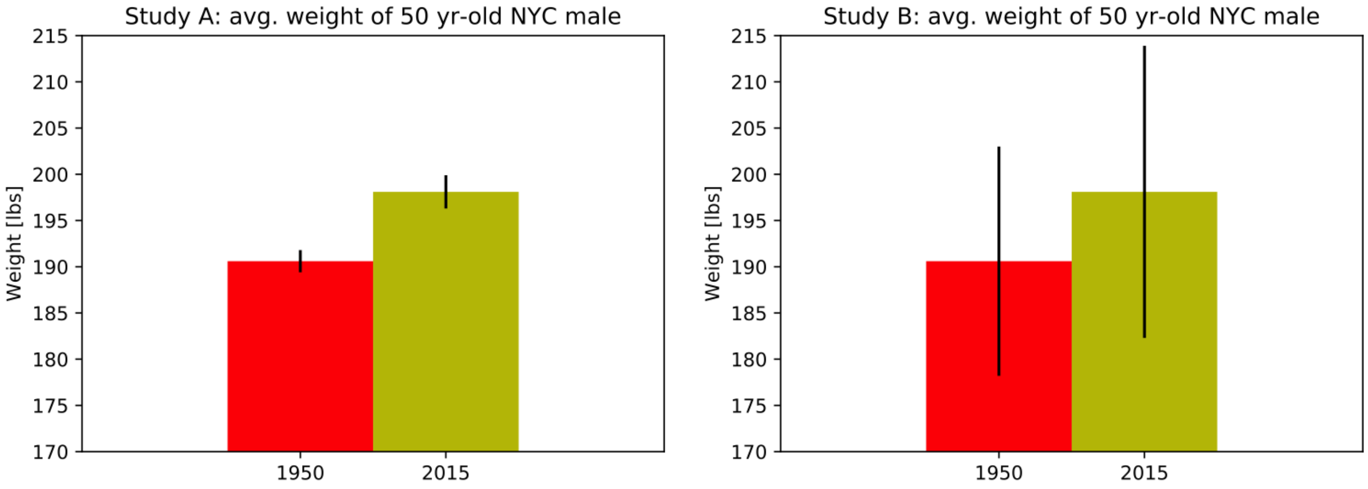
Confidence Intervals

What can you conclude about the following figure?



Douglas Rubin

Now what can you conclude?



- In most situations when presenting data, error bars are absolutely necessary to quantify our uncertainy, so that we can make sound conclusions.
- Confidence intervals are one way to quantify that uncertainty.

Class Outline

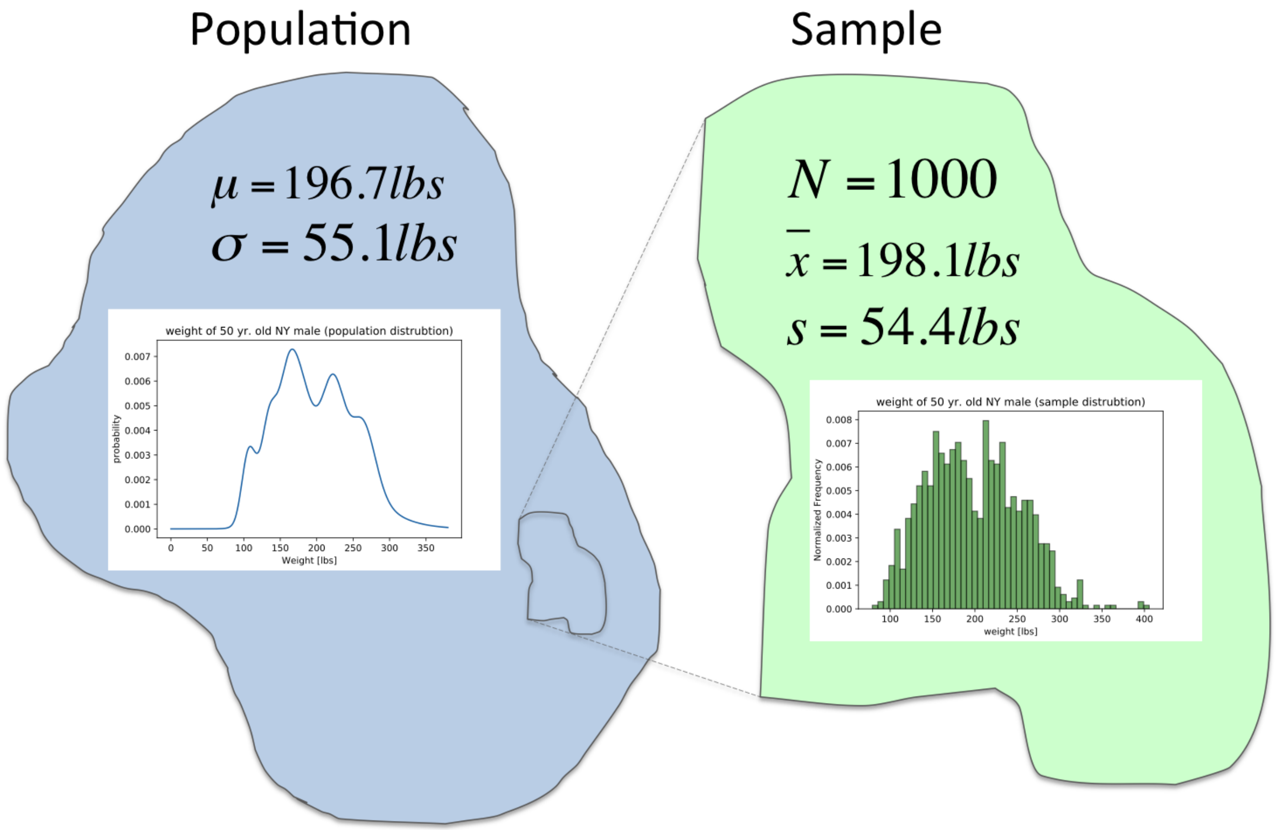
1. Why do we need confidence intervals?
2. What do confidence intervals mean?
   - The Gaussian
   - The Central Limit Theorem
   - confidence intervals
3. How do we calculate confidence intervals?
   - example using a Gaussian distribution
   - example using a t-distribution

I. Why do we need confidence intervals?

Consider a typical sampling experiment:

What is the average (population mean) weight of a 50 year-old New York male? (This will be the ongoing example for the rest of these notes)

Douglas Rubin



μ $\equiv$ population mean, σ $\equiv$ population standard dev., $\bar{x}$ $\equiv$ sample mean, s $\equiv$ sample standard dev.

Measurements vary from sample-to-sample

- We would usually like to infer the population mean value, μ from the sample mean $\bar{x}$.
- However, $\bar{X}$ is a random variable and may be significantly different from the population mean.

Consider $\bar{x}$ (the sample mean weight) calculated from 10 random samples from the 50 year-old male population of NYC (with N = 15):

	1	2	3	4	5	6	7	8	9	10
sample mean	219.09	204.79	192.40	184.06	229.54	174.08	195.18	189.76	178.63	209.05
% error	11.36	4.09	2.21	6.45	16.67	11.52	0.79	3.55	9.21	6.25

Douglas Rubin

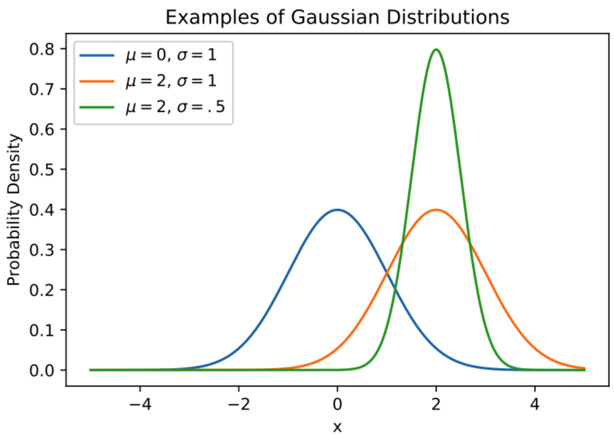
II. What do confidence intervals mean?

before we answer this question, we will to briefly talk about the Gaussian distribution and the Central Limit Thereom

The Gaussian Distribution

Probability density function: $P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{2\sigma^2}}$

Douglas Rubin



Remarks about the Gaussian

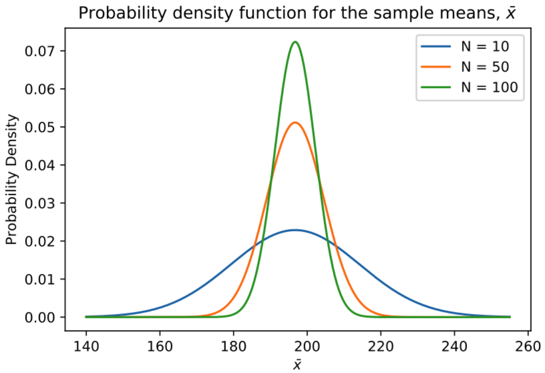
- solely parameterized by its mean, μ, and its standard deviation, σ
- one of the most important distributions in statistics
- pops up everywhere in math and physics (e.g., diffusion/brownian motion, wave functions in quantum mechanics...)
- also known as the Normal distribution

The Central Limit Theorem

Given a population with mean μ and variance σ^2, and a sample of size N, the distribution of the sample mean, $\bar{X}$, converges to a normal distribution as $N \rightarrow \infty$ with

Douglas Rubin

$\mu_{\bar{x}} = \mu$, and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$.



Remarks about the Central Limit Theorem

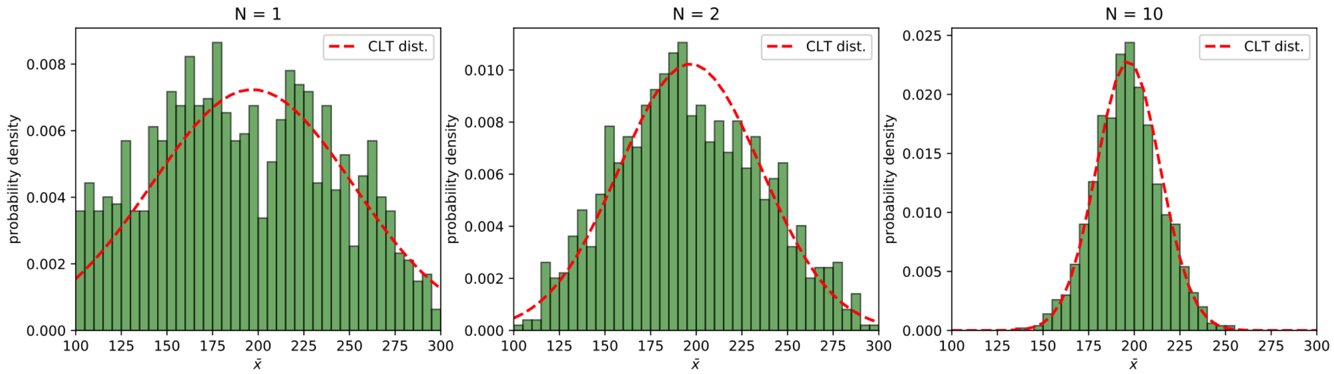
- one of the most important results of statistics
- holds REGARDLESS of what the actual population distribution is (can be any distribution)
- distribution of the sample mean usually converges to a normal distribution reasonably (N ~ 10-100) quickly for most population distributions (i.e., N doesn't have to be huge for the CLT to hold)

The CLT in action

To see the convergence of the sample mean distribution to a Gaussian for our data we may:

Douglas Rubin

1. randomly draw a sample of size N from the population
2. calculate the sample mean
3. repeat 1-2 many times to build up a frequency distribution of the sample mean



Another cool demonstration of the CLT in action is the so called "Bean Machine" (https://www.youtube.com/watch?v=p65aYYuAz-s).

Technically the bean machine draws random values according to a binomial distribution, so that after many draws the frequency distribution is approximately binomial. But, in the limit of many rows of pegs, because of the CLT, it should build up a Gaussian. Why this is is a little technical. See me later if you would like more explanation as to why this machine results in a Gaussian distribution.

Confidence Intervals

- the CLT and the data allow us to construct an interval that, with high probability, we believe bounds the true population mean, μ

Douglas Rubin

- specifically, for large N, for a given value of z_π, the (random) interval

$$\left[\bar{x} - z_\pi \frac{\sigma}{\sqrt{N}},\ \bar{x} + z_\pi \frac{\sigma}{\sqrt{N}}\right]$$

encompasses μ with probability π, where the relationship between π and z_π is given in the following table:

pi	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.999
z_pi	1.44	1.48	1.51	1.55	1.60	1.64	1.70	1.75	1.81	1.88	1.96	2.05	2.17	2.33	2.58	3.290

Remarks

- a note on language: if the value of π that we choose is, say, 0.9, we would say that the computed CI is the "90% confidence interval"

- $\pi = 0.9$ or $\pi = 0.95$ are common values to use to construct a CI

- to construct the interval, σ is usually unknown and typically estimated with s

Proof of previous slide (optional)

The proof of the previous slide is relatively straightforward:

Douglas Rubin

$$
\begin{aligned}
\pi &= \Pr\left(\bar{X} - z_\pi \frac{\sigma}{\sqrt{N}} \le \mu \le \bar{X} + z_\pi \frac{\sigma}{\sqrt{N}}\right) \\
&= \Pr(\bar{X} - z_\pi \sigma_{\bar{x}} \le \mu \le \bar{X} + z_\pi \sigma_{\bar{x}}) \\
&= \Pr(\mu - z_\pi \sigma_{\bar{x}} \le \bar{X} \le \mu + z_\pi \sigma_{\bar{x}}) \\
&= \int_{\mu - z_\pi \sigma_{\bar{x}}}^{\mu + z_\pi \sigma_{\bar{x}}} \frac{1}{\sqrt{2\pi}\sigma_{\bar{x}}} e^{-\frac{1}{2}\left(\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\right)^2} d\bar{X} \\
&= \int_{-z_\pi}^{z_\pi} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz,
\end{aligned}
$$

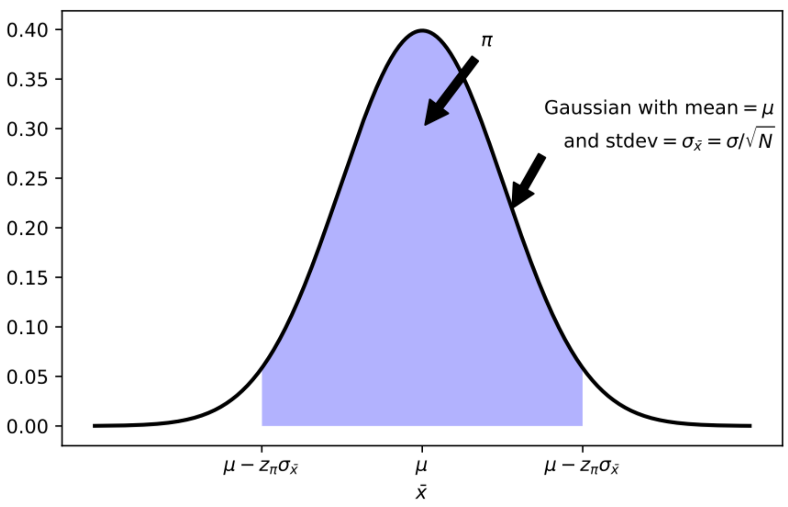
where, in the second and fourth lines I have invoked the CLT, and in the fifth line I have used a change of variables for the integration with $z \equiv (\bar{X} - \mu)/\sigma_{\bar{x}}$.

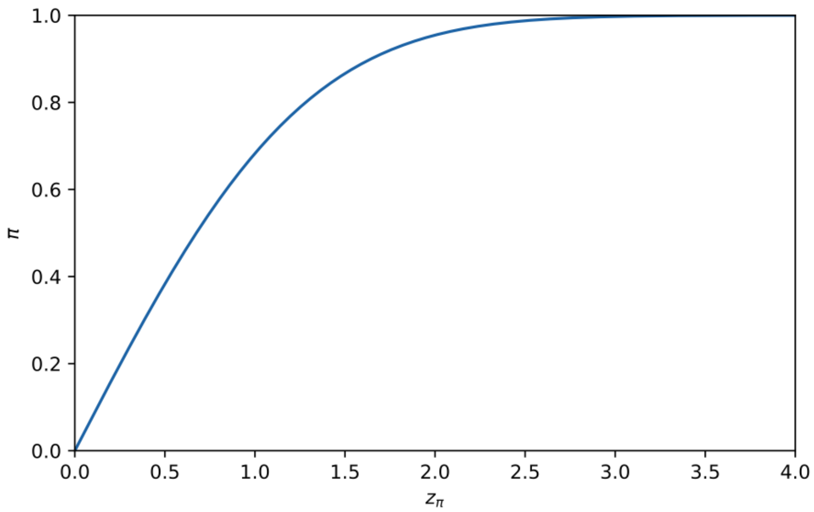
Proof of previous slide (optional) (cont'd)

Douglas Rubin

- From the derivation we see that π equivalently represents the probability that $\bar{x}$ is within z_π standard deviations of the mean.

- The relationship between π and z_π must be calculated using the equation above, and can be easily computed using the so-called error function: $\pi(z_\pi) = \mathrm{erf}\left(\frac{z_\pi}{\sqrt{2}}\right)$.



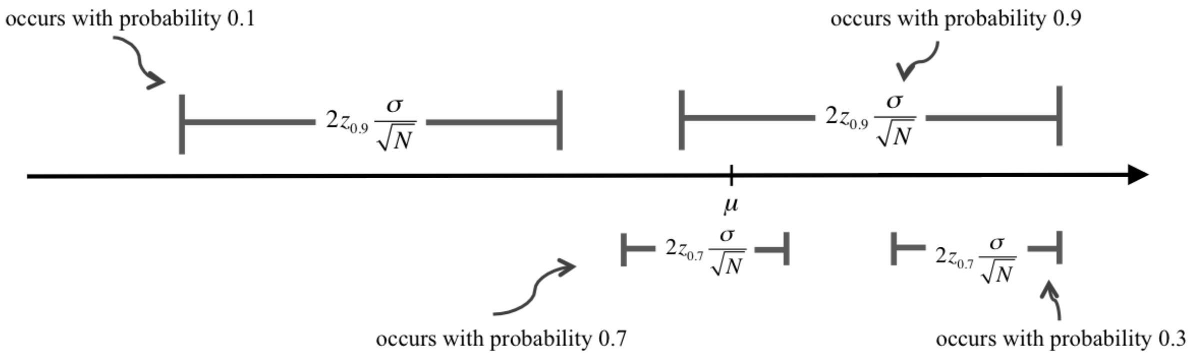


What a CI really means

What does "π is the probability that the random interval $[\bar{x} - z_\pi \frac{\sigma}{\sqrt{N}},\ \bar{x} + z_\pi \frac{\sigma}{\sqrt{N}}]$ encompasses μ" mean?

Douglas Rubin

- Recall the sample mean $\bar{x}$ is a random variable (it randomly changes depending on the exact sample from the population), and thus the associated CI is also random.
- Frequentist POV: if we drew many different samples from the population and and computed their associated CIs at say $\pi = 0.9$, then we would expect about 90% of the intervals to encompass the population mean, μ, and about 10% to not.



- the longer the length of the interval, the higher the probability that the interval encompasses μ (which is why π increases with z_π)

What a CI really means (cont'd)

For example:

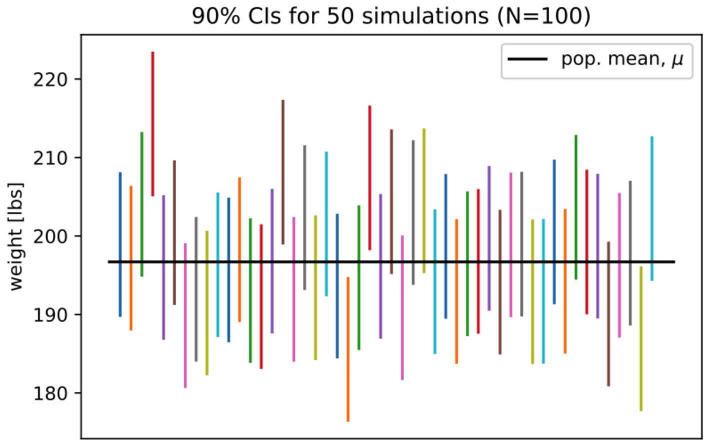
Douglas Rubin

- For 50 realizations of samples from our weight data (all with N = 100), at the 90% confidence level, we would expect about 5 intervals to not contain μ

$$z_{0.9} = 1.64, \frac{\sigma}{\sqrt{N}} = \frac{55.1}{\sqrt{100}} = 5.51$$

$$\Longrightarrow$$

$$CI_{0.9} = \left[\bar{x} - z_{0.9}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{0.9}\frac{\sigma}{\sqrt{N}}\right] = [\bar{x} - 9.04,\ \bar{x} + 9.04]$$



III. How do we calculate confidence intervals?

Example

What is the 95% CI for the "Typical Sampling Experiment" slide?

Douglas Rubin

$N = 1000$

$\bar{x} = 198.1 lbs$

$s = 54.4 lbs$

$z_{0.95} = 1.96$

$\implies$

$$\begin{aligned}
CI_{0.95} &= \left[\bar{x} - z_{0.95}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{0.95}\frac{\sigma}{\sqrt{N}}\right] \\
&\approx \left[\bar{x} - z_{0.95}\frac{s}{\sqrt{N}},\ \bar{x} + z_{0.95}\frac{s}{\sqrt{N}}\right] \\
&= \left[198.1 lbs - 1.96\left(\frac{54.4 lbs}{\sqrt{1000}}\right),\ 198.1 lbs + 1.96\left(\frac{54.4 lbs}{\sqrt{1000}}\right)\right] \\
&= [194.7 lbs,\ 201.5 lbs]
\end{aligned}$$

- this indeed bounds the true population mean of $\mu = 196.7 lbs$
- note that s should be the "unbiased" estimator of the standard deviation

Confidence intervals for a small sample size

Douglas Rubin

- approximating σ with s is only acceptable when N is large
- for small N, the appropriate confidence interval to use is:

$$CI_\pi = \left[\bar{x} - t_\pi \frac{s}{\sqrt{N}},\ \bar{x} + t_\pi \frac{s}{\sqrt{N}}\right],$$

where the t_π values (see next slide) are calculated with a t-distribution instead of a Gaussian

- when using a t-dsitribution, one must specify the so-called degrees of freedom (df) which is simply given by $df = N - 1$
- In the limit that N is large a t-distribution approaches a Gaussian, so why not always use a t-distribution instead of a Gaussian?
    - a Gaussian is a much more well known distribution, and is typically much easier to work with mathematically
- a widely used rule of thumb is that for N less than about 30, one should use a t-distribution

Values of t_π

Douglas Rubin

df	t_0.85	t_0.86	t_0.87	t_0.88	t_0.89	t_0.9	t_0.91	t_0.92	t_0.93	t_0.94	t_0.95	t_0.96	t_0.97	t_0.98	t_0.99	t_0.999
2	2.282	2.383	2.495	2.620	2.760	2.920	3.104	3.320	3.578	3.896	4.303	4.849	5.643	6.965	9.925	31.599
3	1.924	1.995	2.072	2.156	2.249	2.353	2.471	2.605	2.763	2.951	3.182	3.482	3.896	4.541	5.841	12.924
4	1.778	1.838	1.902	1.971	2.048	2.132	2.226	2.333	2.456	2.601	2.776	2.999	3.298	3.747	4.604	8.610
5	1.699	1.753	1.810	1.873	1.941	2.015	2.098	2.191	2.297	2.422	2.571	2.757	3.003	3.365	4.032	6.869
6	1.650	1.700	1.754	1.812	1.874	1.943	2.019	2.104	2.201	2.313	2.447	2.612	2.829	3.143	3.707	5.959
7	1.617	1.664	1.715	1.770	1.830	1.895	1.966	2.046	2.136	2.241	2.365	2.517	2.715	2.998	3.499	5.408
8	1.592	1.638	1.687	1.740	1.797	1.860	1.928	2.004	2.090	2.189	2.306	2.449	2.634	2.896	3.355	5.041
9	1.574	1.619	1.666	1.718	1.773	1.833	1.899	1.973	2.055	2.150	2.262	2.398	2.574	2.821	3.250	4.781
10	1.559	1.603	1.650	1.700	1.754	1.812	1.877	1.948	2.028	2.120	2.228	2.359	2.527	2.764	3.169	4.587
11	1.548	1.591	1.636	1.686	1.738	1.796	1.859	1.928	2.007	2.096	2.201	2.328	2.491	2.718	3.106	4.437
12	1.538	1.580	1.626	1.674	1.726	1.782	1.844	1.912	1.989	2.076	2.179	2.303	2.461	2.681	3.055	4.318
13	1.530	1.572	1.616	1.664	1.715	1.771	1.832	1.899	1.974	2.060	2.160	2.282	2.436	2.650	3.012	4.221
14	1.523	1.565	1.609	1.656	1.706	1.761	1.821	1.887	1.962	2.046	2.145	2.264	2.415	2.624	2.977	4.140
15	1.517	1.558	1.602	1.649	1.699	1.753	1.812	1.878	1.951	2.034	2.131	2.249	2.397	2.602	2.947	4.073
16	1.512	1.553	1.596	1.642	1.692	1.746	1.805	1.869	1.942	2.024	2.120	2.235	2.382	2.583	2.921	4.015
17	1.508	1.548	1.591	1.637	1.686	1.740	1.798	1.862	1.934	2.015	2.110	2.224	2.368	2.567	2.898	3.965
18	1.504	1.544	1.587	1.632	1.681	1.734	1.792	1.855	1.926	2.007	2.101	2.214	2.356	2.552	2.878	3.922
19	1.500	1.540	1.583	1.628	1.677	1.729	1.786	1.850	1.920	2.000	2.093	2.205	2.346	2.539	2.861	3.883
20	1.497	1.537	1.579	1.624	1.672	1.725	1.782	1.844	1.914	1.994	2.086	2.197	2.336	2.528	2.845	3.850

Example for small N

You are given the following randomly collected weight data for 50 year-old NYC males:
Douglas Rubin
198.0lbs, 151.0lbs, 208.8lbs, 234.7lbs, 144.0lbs. What is the 95% confidence interval for this dataset?

$N = 5$

$df = 5 - 1 = 4$

$t_{0.95} = 2.78$

$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 187.3lbs$

$s = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(x_i - \bar{x})^2} = 38.8lbs$

$\Longrightarrow$

$$\begin{aligned} CI_{0.95} &= \left[\bar{x} - t_{0.95}\frac{s}{\sqrt{N}},\ \bar{x} + t_{0.95}\frac{s}{\sqrt{N}}\right] \\ &= \left[187.3lbs - 2.78\left(\frac{38.8lbs}{\sqrt{5}}\right),\ 187.3lbs + 2.78\left(\frac{38.8lbs}{\sqrt{5}}\right)\right] \\ &= [139.1lbs,\ 235.5lbs] \end{aligned}$$

Some final remarks on CIs/error bars

Douglas Rubin

- To make valid assertions from data you should almost always quantify uncertainty and represent that uncertainty with error bars
- When used as error bars in plots, typically the 90 or 95% CI levels are used (the level used should be indicated in the figure)

- Not all error bars represent CIs. Other types include:

    - standard errors of an estimator (eg: linear regression)

      (note that $\sigma/\sqrt{N}$ is the SE of the estimator $\bar{x}$)

    - credible regions from a Bayesian posterior (eg: Bayesian linear regression)

- When looking at a plot, knowing which types of uncertainty the error bars represent is crucial for proper interpretation.

- When creating a plot, knowing which type of error bars to include is crucial to get your point across convincingly.