Lecture 8: Variance in MLR and Inference

Pierre Biscaye

Fall 2022

R^2 in multiple regression

- We can still calculate R^2 as a measure of goodness-of-fit

$$R^2 = 1 - \frac{SSR}{SST_y} = 1 - \frac{\sum \hat{u_i^2}}{\sum(y_i - \bar{y})^2} \tag{1}$$

- R^2 can only increase when you add more variables
- R outputs the "Adjusted R^2" with regression results which corrects for this

$$\bar{R}^2 = 1 - (\frac{SSR}{SST_y})\frac{n-1}{n-k-1} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \tag{2}$$

Adjusted R^2

$$\bar{R}^2 = 1 - (\frac{SSR}{SST_y})\frac{n-1}{n-k-1} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

- n is sample size and k is number of explanatory variables ($n-k-1$ is your *degrees of freedom*)
- Adjusted R^2 penalizes you for adding more variables to the model, particularly if they have limited (additional) explanatory power
  - R^2 can never decrease when adding more variables, but $\bar{R}^2$ can
- Will primarily consider $\bar{R}^2$ when comparing models going forward

To Jupyter!

Variability in MLR estimators

- We need concept of R^2 in MLR setting to calculate variance of $\hat{\beta}$s
  - Matters for estimating range in which true βs are likely to fall
- MLR5 (Homoskedasticity): $var(u|x_1, ..., x_k) = \sigma_u^2$
  - The variance of the error term is unrelated to all of the model x variables.
- Under MLR1-MLR5:

$$var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)} = \frac{\sigma_u^2}{\sum_i (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)} \tag{3}$$

$$\tag{4}$$

- R_j^2 is the (unadjusted) R^2 from a regression of x_j on $x_1, ..., x_{j-1}, x_{j+1}, ..., x_k$

Under MLR1-MLR5, we have an expression for $\widehat{var(\hat{\beta}_j)}$

- Still don't observe σ_u^2

$$\hat{\sigma_u^2} = \frac{\sum_i \hat{u_i}^2}{n-k-1} \tag{5}$$

- $n-k-1$ is the *degrees of freedom*

$$\widehat{var(\hat{\beta}_j)} = \frac{\sum_i \hat{u_i}^2}{(n-k-1)\sum_i(x_{ji} - \bar{x_j}^2)(1-R_j^2)} \tag{6}$$

- Compare to SLR

$$\widehat{var(\hat{\beta_1})} = \frac{\sum_i \hat{u_i}^2}{(n-2)\sum_i(x_{1i} - \bar{x_1}^2)} \tag{7}$$

MLR $\hat{\beta}$ standard error

$$SE(\hat{\beta}_j) = \sqrt{\widehat{var(\hat{\beta}_1)}} \tag{8}$$

$$= \sqrt{\frac{\sum_i \hat{u}_i{}^2}{(n-k-1)\sum_i (x_{ji} - \bar{x}_j{}^2)(1-R_j^2)}} \tag{9}$$

- is smaller when $\sum_i \hat{u}_i{}^2$ is smaller
- is smaller when we have more degrees of freedom $(n-k-1)$
- is smaller when there is more variance in x_j
- is smaller when R_j^2 is smaller

R_j^2

- Remember "partialing out" in MLR.
  - We are only using unexplained variation in x_j to estimate $\hat{\beta}_j$; need to account for this in estimating its variance
- Consider perfect multicollinearity: suppose $x_j = \delta_0 + \delta 1 x_1 + ... + \delta_{j-1}x_{j-1} + \delta_{j+1}x_{j+1} + ... + \delta_k x_k$
- MLR3 fails, $R_j^2 = 1$, and $var(\hat{\beta}_j) \to \infty$
- With near perfect multicollinearity: $x_j = \delta_0 + \delta 1 x_1 + ... + \delta_{j-1}x_{j-1} + \delta_{j+1}x_{j+1} + ... + \delta_k x_k + u$ but u are small
- Then $\hat{\beta}_j$ can be estimated, but is quite variable (R_j^2 is close to 1)
  - Hard to isolate precise effect of x_j when it is closely related to other variable we want to hold constant

To Jupyter!

Inference and Confidence Intervals

- We now have a means of estimating $\hat{\beta}$ and we know how variable it will be (across different sample draws)
- For a given estimate $\hat{\beta}$, what do we learn about the true parameter β?
  - Can we rule out the hypothesis that β takes on some specific values?
  - This is the objective of *inference*

Focus on a simpler estimator

- We have a random variable X with $E[X] = \mu$ and $var(X) = \sigma^2$.
- We draw sample of n observations of X : $x_1, x_2, ..., x_n$
- Consider $\bar{X}$. In Lecture 5, we showed that for a random sample $\bar{X}$ is an unbiased (consistent) estimator for μ, and that we can think of $\bar{X}$ as a random variable with its own variance

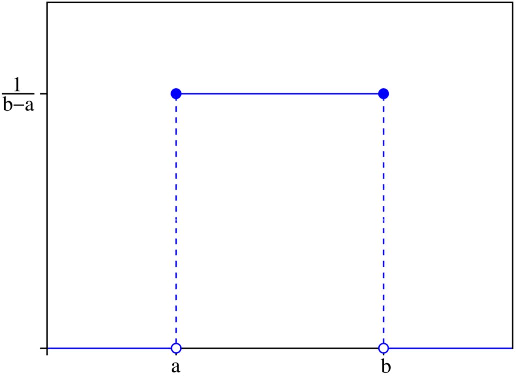
$$E[\bar{X}] = E[\frac{1}{n}\sum_i X_i] = \frac{1}{n}\sum_i E[X_i] = \frac{1}{n} n\mu = \mu \tag{10}$$

$$Var(\bar{X}) = var(\frac{1}{n}\sum_i X_i) = \frac{1}{n^2}\sum_i var(X_i) = \frac{n\sigma_x^2}{n^2} = \frac{\sigma_x^2}{n} \tag{11}$$

Inference with $\bar{X}$

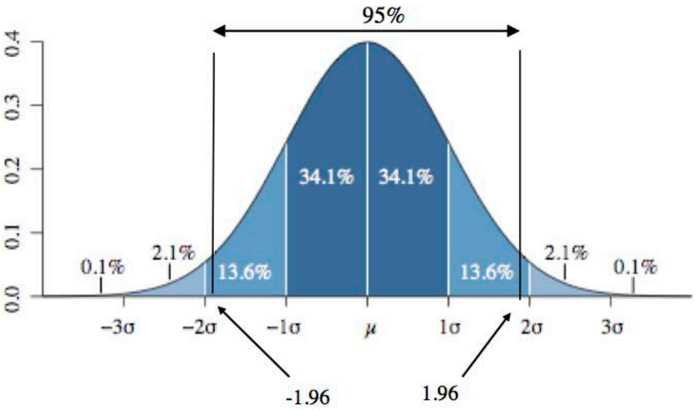
- If the true population mean is μ_0, what is the probability that I observe a specific value of $\bar{X}$ for a particular sample?
- If the probability is low enough, we will reject the hypothesis that the true mean is μ_0
- To know this, need to know how likely different values of $\bar{X}$ are when the mean is μ_0.
- Expect that values of $\bar{X}$ that are close to μ_0 are more likely if μ_0 is the true mean... but how much more likely?
- Mean and variance of $\bar{X}$ are not sufficient to know this. Need to know the *distribution* of $\bar{X}$.

For example, a uniform random variable



- If a random variable X is distributed $U(a, b)$ then the probability if falls in any interval in a, b is the same.
- This would imply that being far from the mean $((b - a)/2)$ is not less likely than being close to the mean.

Or, a normal random variable



- 64.2% probability of being within 1 standard deviation of the mean
- 0.2% probability of being more than 3 standard deviations from the mean

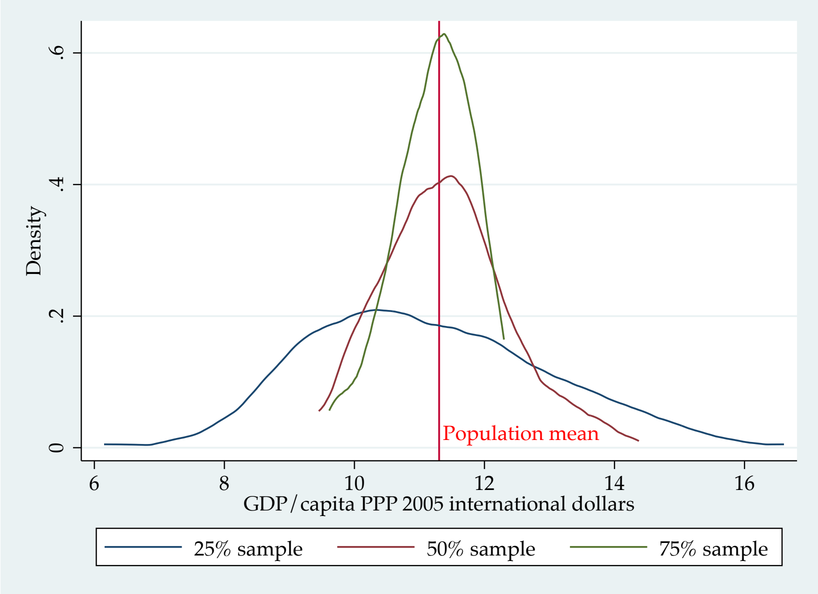
Central Limit Theorem

- Fortunately, an important theorem tells us exactly how $\bar{X}$ will be distributed.
- Suppose we have a random sample of observations $X_1, X_2, ..., X_n$ of a random variable X with true mean μ_x and variance σ_x^2. Then for n sufficiently large

$$\bar{X} \sim\approx N(\mu_x, \frac{\sigma_x^2}{n}) \tag{12}$$

- $\bar{X}$ is approximately *normally distributed* with mean μ and variance σ_x^2 / n
- This is true no matter what the underlying distribution of X is

Recall from L5: distribution of mean GDP/capita estimates



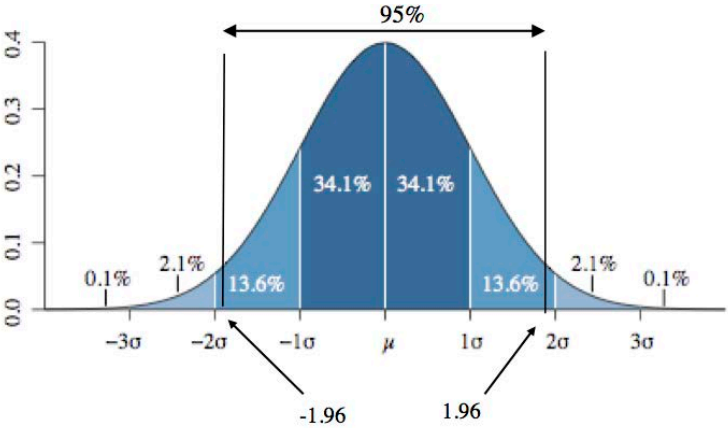
Properties of Normal Random Variables

- Normal Random Variables are defined by their mean and variance
- If $Y \sim N(\mu_Y, \sigma_Y^2)$
- Then $Z = aY + b$ has:
  - mean $a\mu_Y + b$
  - variance $a^2\sigma_Y^2$
  - and distribution $N(a\mu_Y + b, a^2\sigma_Y^2)$

Usefulness of Normal properties

- Consider $Z = \frac{\bar{X}-\mu_x}{\sigma_x/\sqrt{n}} = \frac{1}{\sigma_x/\sqrt{n}}\bar{X} - \frac{\mu_x}{\sigma_x/\sqrt{n}}$ (a 'normalization' of $\bar{X}$)
- Since $\bar{X} \sim N(\mu, \frac{\sigma_x^2}{n})$, we know that
  $Z \sim N(\frac{\mu_x}{\frac{\sigma_x}{\sqrt{n}}} - \frac{\mu_x}{\frac{\sigma_x}{\sqrt{n}}}, \frac{\frac{\sigma_x^2}{n}}{\frac{\sigma_x^2}{n}}) = N(0,1)$
- So we can always transform a normal random variable into *standard normal* random variable: can use properties of that distribution to conduct inference on a wide variety of random variables.

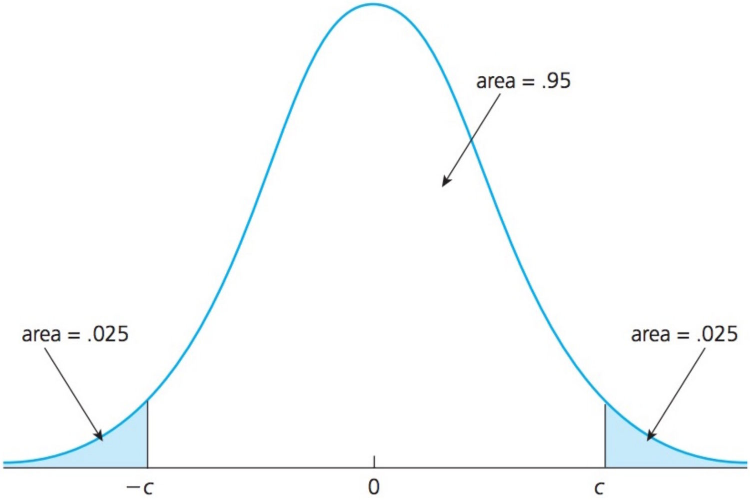
Normal distribution



If Z is standard Normal

- Recall units for normalized variables are standard deviations away from the mean
- $Pr(-1.96 > Z) = Pr(Z > 1.96) \approx 0.025$
- $Pr(-1.96 < Z < 1.96) \approx 0.95$
- $Pr(-1.67 < Z < 1.67) \approx 0.90$
- $Pr(-2.56 < Z < 2.56) \approx 0.99$
- Values of Z associated with particular probabilities are called *critical values*

Finding critical values



Putting this together

- If the true mean of X is μ_X and the variance is σ_X^2
- Then $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$, and

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1) \tag{13}$$

- This allows us to develop *Confidence Intervals*
  - Use what we know about this distribution to calculate critical values for a given probability threshold (confidence level)
  - Use this to back out range of values (confidence interval) of $\bar{X}$ that contain the true μ_X at this confidence level

Confidence intervals

$$Pr(-1.96 < Z < 1.96) = 0.95 \tag{14}$$

$$Pr(-1.96 < (\frac{\bar{X} - \mu_X}{\sigma_x / \sqrt{n}}) < 1.96) = 0.95 \tag{15}$$

$$Pr(-1.96 * \frac{\sigma_X}{\sqrt{n}} < \bar{X} - \mu_X < 1.96 * \frac{\sigma_X}{\sqrt{n}}) = 0.95 \tag{16}$$

$$Pr(-\bar{X} - 1.96 * \frac{\sigma_X}{\sqrt{n}} < -\mu_X < -\bar{X} + 1.96 * \frac{\sigma_X}{\sqrt{n}}) = 0.95 \tag{17}$$

$$Pr(\bar{X} - 1.96 * \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + 1.96 * \frac{\sigma_X}{\sqrt{n}}) = 0.95 \tag{18}$$

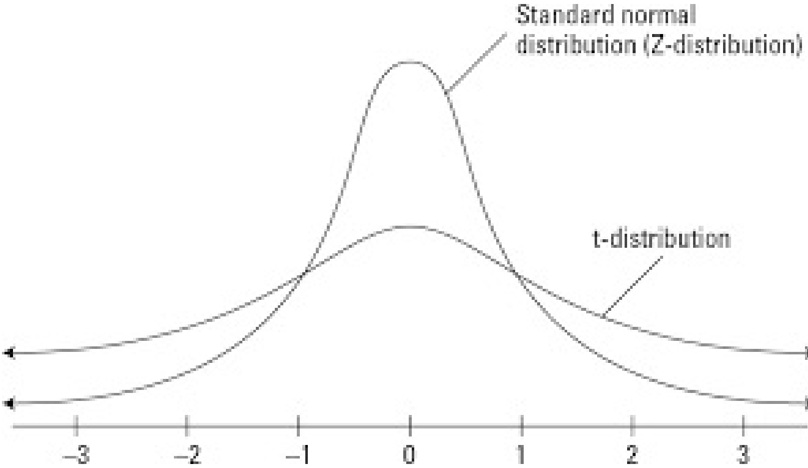
Confidence Intervals

- We use $Pr(\bar{X} - 1.96 * \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + 1.96 * \frac{\sigma_X}{\sqrt{n}}) = 0.95$ to develop a 95% *Confidence Interval* for μ given an estimate of $\bar{X}$
- Interval will be small if n is large and if σ_X^2 is small
- We know that with 95% probability, our confidence interval (CI) contains the true value μ
- Important concept: the probability is in the CI we construct (which varies with our estimate of $\bar{X}$), not μ. μ is either in the CI or it is not.
- So we can always estimate an interval that has a 95% chance of containing μ.

s_x^2 vs σ_x^2

- Issue: we don't know σ_x^2
- What if we use $s_x^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$?
- We can construct $t = \frac{\bar{X} - \mu}{s_X / \sqrt{n}}$; lose Normal distribution
- t is distributed t with $n-1$ degrees of freedom
- Since t is distributed t_{n-1} (and not standard normal) need different critical values, but process is the same

t and normal distributions



Critical values of t distributions

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Example: student height

- Suppose I sampled 41 students at random from the class and measured their heigh (X)
- I want a 95% Confidence Interval for the mean height (μ_X)
- In our sample, $\bar{X} = 67$ inches, and $s_X^2 = 4$

$$Pr(-c_{0.025} < \frac{\bar{X} - \mu}{s_X / \sqrt{n}} < c_{0.025}) = 0.95 \tag{19}$$

$$Pr(-c_{0.025} < \frac{67 - \mu}{2 / \sqrt{41}} < c_{0.025}) = 0.95 \tag{20}$$

Finding $c_{0.025}$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Finding our CI

$$Pr(-2.021 < \frac{67-\mu}{2/\sqrt{41}} < 2.021) = 0.95 \quad (21)$$

$$Pr(67 - 2.021 * (2/\sqrt{41}) < \mu < 67 + 2.021 * (2/\sqrt{41})) = 0.95 \quad (22)$$

$$Pr(67 - 2.021 * 0.31 < \mu < 67 + 2.021 * .31) = 0.95 \quad (23)$$

$$Pr(66.37 < \mu < 67.63) = 0.95 \quad (24)$$

- Our estimated Confidence Interval is (66.37,67.63)
- Interpretation:
  - With 95% probability, the true mean height in the class is between 66.37 and 67.63 inches? Not quite.
  - The CI for a given sample mean is fixed, so either μ is in it or it isn't.
  - The interpretation is that for 95% of samples, μ will be inside the calculated interval.

What if n was smaller?

- Suppose I had only sampled 10 students, but still had $\bar{X} = 67$ inches, and $s_X^2 = 4$
- This changes (increases) the critical value, and also changes (increases) the estimated variance of the sample mean

$$Pr(-c_{0.025} < \frac{\bar{X} - \mu}{s_X / \sqrt{n}} < c_{0.025}) = 0.95 \tag{25}$$

$$Pr(-c_{0.025} < \frac{67 - \mu}{2/\sqrt{10}} < c_{0.025}) = 0.95 \tag{26}$$

$$Pr(-2.228 < \frac{67 - \mu}{2/\sqrt{10}} < 2.228) = 0.95 \tag{27}$$

$$Pr(65.59 < \mu < 68.41) = 0.95 \tag{28}$$

Formalizing our understanding of confidence intervals

- Let's run some simulations to help with our understanding of how to interpret confidence intervals

To Jupyter!

- To recap: With a 95% confidence level, μ will be inside the calculated CI for 95% of samples.
- For any individual sample, the CI is an estimate of the 95% probability range and μ will either be inside or outside the estimated CI.
- This estimated CI is useful for identifying likely values of μ, particularly if the interval is small.