Mathematics 231

Lecture 21

Liam O'Brien

Announcements

- Reading
  - Today

Today	M&M 6.1	349-369
	M&M 7.1	418-422
	M&M 8.1	488-493
Next class	M&M 6.2	372-390

Topics

- T-distribution
- Confidence intervals for μ with unknown σ
- Confidence intervals for p

What if We Don't Know σ?

- We usually don't know the population standard deviation, σ.
- Estimate σ using the sample standard deviation s.
- This will change our formula for the confidence interval since we will no longer be under a normal distribution (the CLT doesn't help us here).

Confidence Intervals (σ unknown)

We can estimate the sd of $\bar{x}$ using $\dfrac{s}{\sqrt{n}}$;

this is called the standard error of $\bar{x}$.

Estimating σ with s introduces a new source of variation. As a result the CI needs to be wider. How much wider?

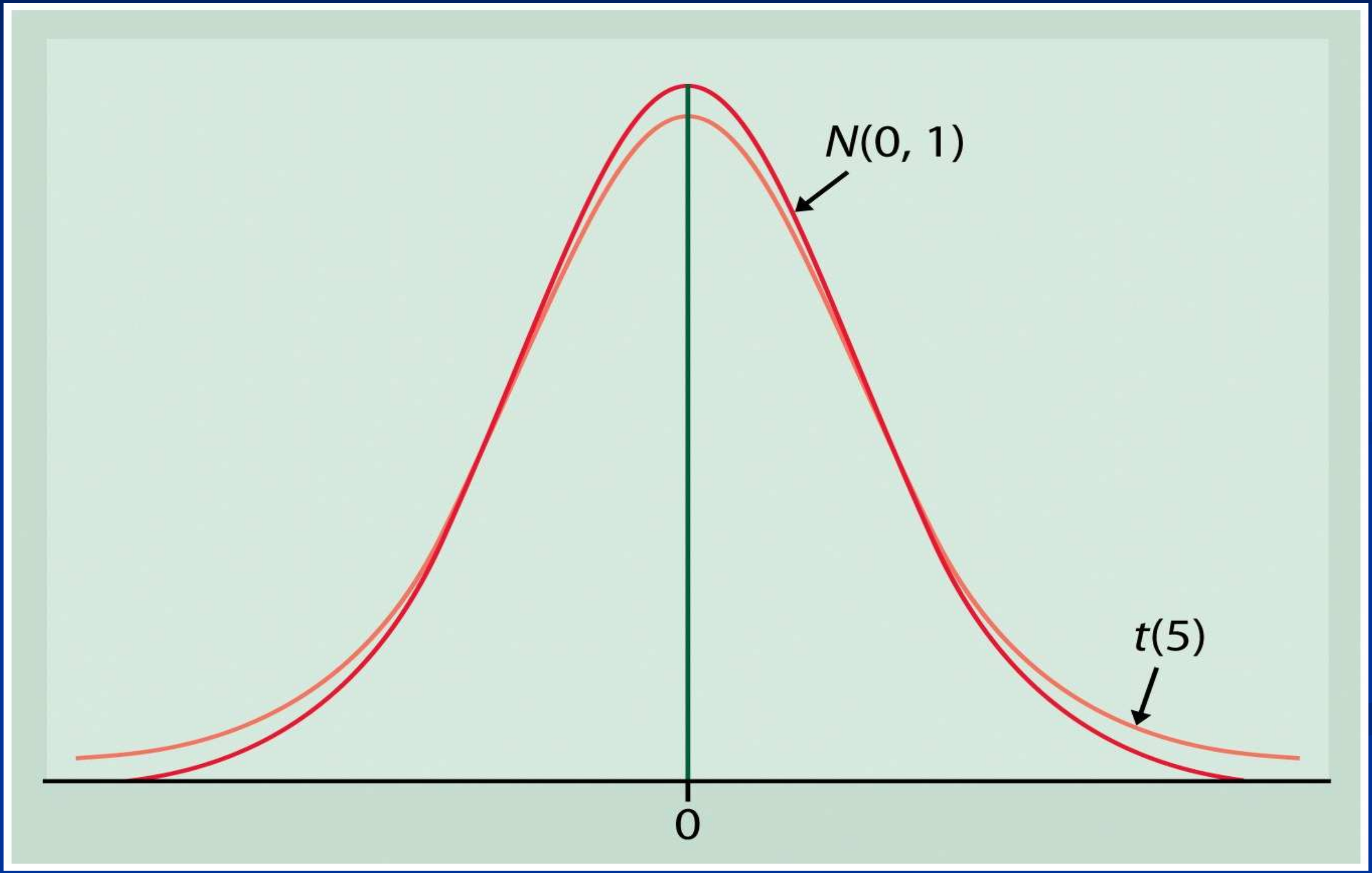
Confidence Intervals (σ unknown)

- When n is small, s doesn't estimate σ well, so the CI needs to be quite a bit wider.
- When n is large, s estimates σ better, and the CI only needs to be slightly wider.
- How do we make the interval wider?
- Make the multiplier z* larger; use the t-distribution (t*) instead of the standard normal distribution.

Student's t-distribution

- Properties of the t-distribution
  - "Bell-shaped" and symmetric similar to the normal distribution.
  - More spread out than the normal distribution.
  - Exact shape depends on its degrees of freedom.
  - As the number if degrees of freedom increases, the corresponding t-distribution looks more like a standard normal distribution.

Normal versus t-distribution



Confidence Intervals (σ unknown)

Before $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ had a standard normal distribution (CLT).

Now $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom (df)

df = amount of information available in data for estimating σ.

Sampling Distribution of a Sample Mean

- Distribution of values taken by the sample mean in all possible samples of size **n** from the population with unknown σ.
- For sample 1: SRS of size n → $\bar{x}_1, s_1, \frac{\bar{x}_1 - \mu}{s_1/\sqrt{n}}$
- For sample 2: SRS of size n → $\bar{x}_2, s_2, \frac{\bar{x}_2 - \mu}{s_2/\sqrt{n}}$
- Etc.

CI's from the t-distribution

- Before, we had a CI for μ given by,

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Now the CI is given by

$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ and $t^* \frac{s}{\sqrt{n}}$ is the margin of error.

Note: This interval is exact when the underlying population has a normal distribution, but is approx. correct when n is "large."

Example: Housing Prices

- In an SRS of size 25, we obtain a sample mean of 215, and a sample sd of of 42.

$$95\%\ \text{CI: } \bar{x} \pm t^{*} \frac{s}{\sqrt{n}}$$

$$= 215 \pm 2.064 \frac{42}{\sqrt{25}}$$

$$= (197.7, 232.4)$$

Note with n = 25, df = n-1, and $t^{*} = 2.064$

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

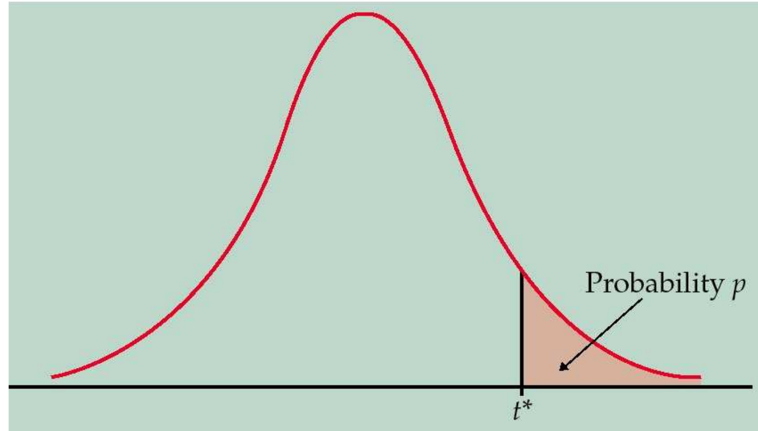


TABLE D *t* **distribution critical values**

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

Confidence level C

Note: For a t-distribution with 24 df, 95% of the area falls between -2.064 and +2.064.

Finding t Multipliers in Stata

- Use the command line for this.
- To find the value that cuts off a certain area (p) to the right under a t-distribution with df degrees of freedom:
- display invttail(df, p)
- For example, "display ttail(24,.025)" gives 2.064.

Sampling Distribution for a Proportion

- Recall: when n is large (np ≥ 10 and n(1-p) ≥ 10) then

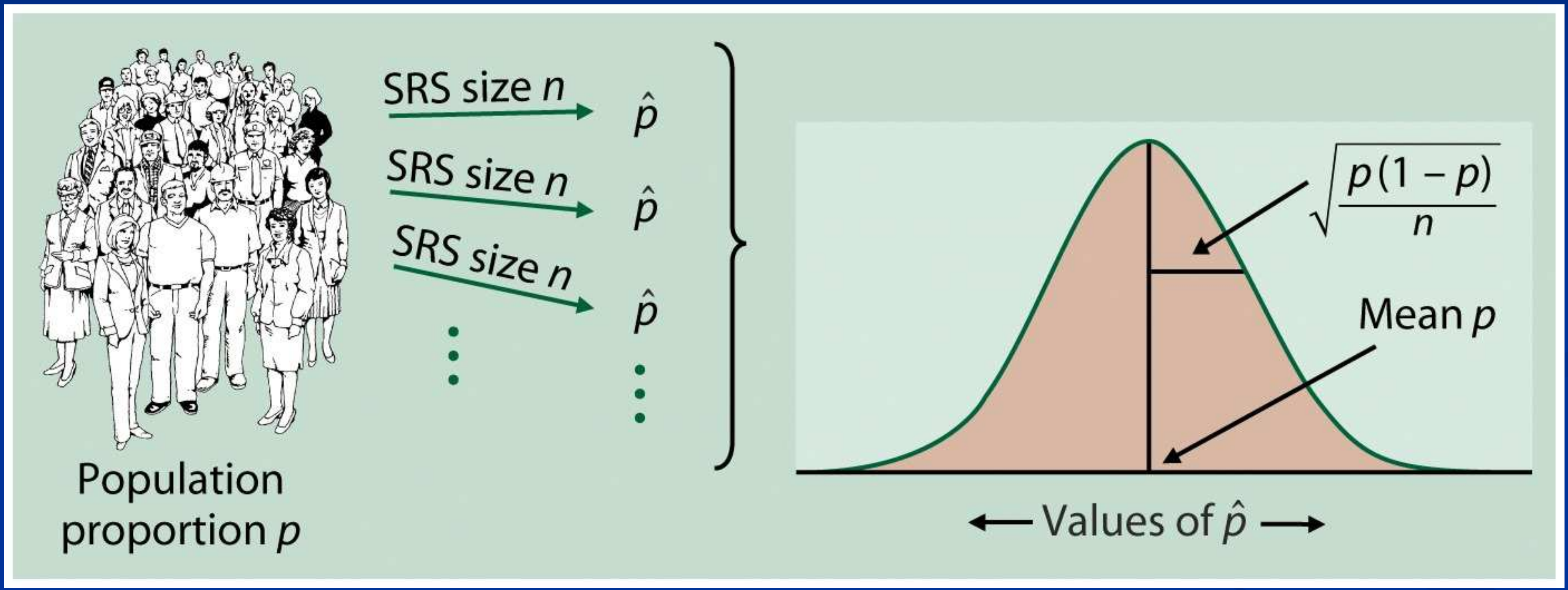
  the sampling distribution of $\hat{p} = \dfrac{x}{n}$

  is approximately normal with

  $$\mu_{\hat{p}} = p$$

  $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sampling Distribution of a Proportion



CI for a Proportion

- A "traditional" 95% CI for p:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = \frac{x}{n}$ is calculated from our sample.

The multiplier can of course be changed for any confidence level.

CI for a Proportion

- There is a problem with this formulation when p is close to 0 or 1.
- **Adjustment:** Pretend we have 4 additional observations, 2 successes and 2 failures.

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

where $\tilde{p} = \dfrac{x+2}{n+4}$ is the adjusted sample proportion.

Example

- In an SRS of 100 adults, obtain x = 95 who want better health insurance.

95% CI for p: $\tilde{p} \pm 1.96\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{n+4}}$

$$\tilde{p} = \frac{x+2}{n+4} = \frac{95+2}{100+4} = 0.933$$

95% CI for p:

$$0.933 \pm 1.96\sqrt{\frac{0.933(1-0.933)}{104}} = (0.885, 0.981)$$

A traditional 95% CI for p is $(0.907, 0.993)$