Monday January 23 Announcements/Reminders

- I will be out-of-town next Monday, so there will be no class meeting or office hour on Monday January 30.
- I will make a video to cover the material for next Monday's class, and it will be available before class time.
- Homework 2 is due this Wednesday. You may submit Homework 1 Problem 4 with Homework 2 for full credit (if you missed it on Hw01).

Plan for today: Review of statistical inference

- The t distribution
- Confidence interval for the mean μ
- Discuss Hw02 Problem 4

Recall: On the previous page, we assumed we have a random sample of size n from a population with expected value μ and standard deviation σ.
Then, the distribution of the sample mean $\bar{X}$ is given in terms of its z-score to be:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

Now we move on to another famous distribution in statistics.
Two practical issues arise in applying the above formula:

1. We don't know the value of the population SD σ
2. We need a known distribution for all sample sizes, not just large samples.

In 1908, the paper "The probable error of a mean, " by Student, was published in the journal *Biometrika*. The name "Student" was a pseudonym for William Sealy Gossett, who was a statistician for Guinness Brewery and was required to use a pseudonym.

In this paper, the t distribution was derived for the quantity $\frac{\bar{X}-\mu}{S/\sqrt{n}}$, where S is the sample SD.

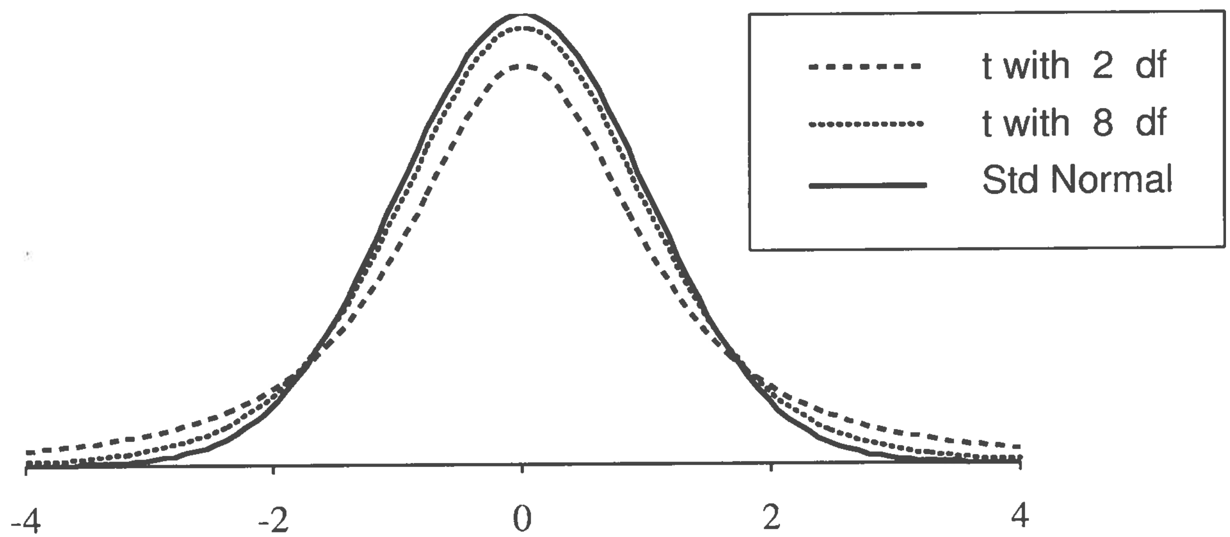
The t distribution

When we substitute the sample standard deviation S for σ, the distribution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$ is no longer a normal distribution.

Fact If we have a random sample of size n from a normal distribution with mean μ and standard deviation σ, then

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{("}t\text{ with } n-1 \text{ degrees of freedom")}$$

The t Distribution



df = 5 $t_{.025} = 2.571$
df = 10 $t_{.025} = 2.228$
df = 20 $t_{.025} = 2.086$
df = 30 $t_{.025} = 2.042$
df = ∞ $t_{.025} = 1.96$

$t_{.025}$ I prefer to label

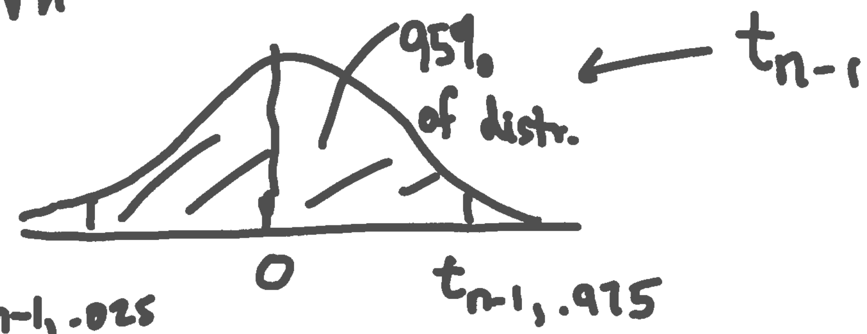
$t_{.975}$ ← this way

Note: $t_{.025} = -t_{.975}$ by symmetry of the curve.

85%

Derive the confidence interval for μ based on the t distribution. Start from the fact on p. 22

$$\frac{\bar{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$



$$P\left\{t_{.025} < \frac{\bar{Y}-\mu}{s/\sqrt{n}} < t_{.975}\right\} = .95$$

By algebra, re-express the event on the l.h.s.:

$$P\left\{t_{.025}\frac{s}{\sqrt{n}} < \bar{Y}-\mu < t_{.975}\frac{s}{\sqrt{n}}\right\} = .95$$

Next: subtract $\bar{Y}$, mult. by (-1), change direction of inequalities

$$P\left\{\bar{Y}-t_{.975}\frac{s}{\sqrt{n}} < \mu < \bar{Y}+t_{.975}\frac{s}{\sqrt{n}}\right\} = .95$$

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Example: You want to see how much you weigh. Your bathroom has a scale which gives readings which are unbiased (have mean equal to the true weight of the object on the scale) and unknown SD. You weigh yourself three times, getting 149, 151, 150. What is a 95% CI for your true weight? Let μ be your true weight.

Calculate $\bar{Y} = 150$ $\quad S = 1$ $\quad S = \sqrt{\frac{1}{n-1}\sum(Y_i - \bar{Y})^2}$

Now $n = 3$ we need $t_{2,.975} = 4.303$

Find $SD(\bar{Y}) = \frac{s}{\sqrt{n}} = \frac{1}{\sqrt{3}} = .5774$

Now the 95% CI is $\bar{Y} \pm t_{2,.975} \frac{s}{\sqrt{n}}$

$150 \pm 4.303(.5774)$

150 ± 2.4846

(margin of error is 2.4846)

$(147.5, 152.5) \rightarrow$ 95% CI for μ

95% CI for μ: $\bar{Y} \pm t_{.025,2} \frac{S}{\sqrt{3}}$
We get $\bar{Y} = 150$, and $S = 1$. From the t-table, we find $t_{.025,2} = 4.303$.
The 95% CI is $150 \pm 4.303 \times \frac{1}{\sqrt{3}}$, or $(147.5, 152.5)$

Meaning of 95% confidence interval (illustrated in scenario of one-sample, inference about mean μ)

Short interpretation: We are 95% confident that the true parameter μ is within the interval $(147.5, 152.5)$.

Long interpretation: If we repeated the experiment many times, always with sample size $n = 3$, and formed the CI by the same method for each repetition of the experiment, then for about 95% of the experiments, the CI would contain the true mean μ.

Comparison of means from two normal samples (any sample sizes)

Basic Framework

Assume:

1. The Y's are a random sample of size n_Y from a normal population with mean μ_Y and SD σ_Y, both unknown; $\bar{Y}$ is the sample mean and S_Y is the sample SD.
2. The Z's are a random sample of size n_Z from a normal population with mean μ_Z and SD σ_Z, both unknown; $\bar{Z}$ is the sample mean and S_Z is the sample SD.
3. The two samples are independent of one another.
4. The two standard deviations are equal; that is, $\sigma_Y = \sigma_Z$.

Want:

1. A confidence interval for $\mu_Y - \mu_Z$.
2. Hypothesis tests concerning $\mu_Y - \mu_Z$.

HW02 Problem 4 We discussed the following

Two steps:

1) First get expected value & variance of $\bar{Y}$ and $\bar{Z}$ using results for sample means in terms of μ_Y, μ_Z, and σ^2

Note: The two sample sizes are unequal $n_Y = 3$, $n_Z = 5$

2) Then use rules p.18 to get $E(.5\bar{Y} + .5\bar{Z})$ and $Var(.5\bar{Y} + .5\bar{Z})$