General Statistics

Ch En 479
Unit Operations

Quantifying variables
(i.e. answering a question with a number)

1. Directly measure the variable.
   - referred to as “measured” variable

   ex. Temperature measured with thermocouple

2. Calculate variable from “measured” or “tabulated” variables
   - referred to as “calculated” variable

   ex. Flow rate $\dot{m} = \rho A v$ (measured or tabulated)

Each has some error or uncertainty

Outline

1. Error of Measured Variables
2. Comparing Averages of Measured Variables

1. Error of Measured Variable

Several measurements are obtained for a single variable (i.e. T).

Questions

- What is the true value?
- How confident are you?
- Is the value different on different days?

Some definitions:

$\bar{x}$ = sample mean
s = sample standard deviation

μ = exact mean
σ = exact standard deviation

As the sampling becomes larger:

$\bar{x} \rightarrow \mu$ $s \rightarrow \sigma$

t chart z chart

not valid if bias exists
(i.e. calibration is off)

How do you determine the error?

- Let's assume "normal" Gaussian distribution
- For small sampling: s is known
- For large sampling: σ is assumed

$$\bar{x} = \sum_i \frac{x_i}{n}$$

small

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

we'll pursue this approach

Use t tables for this approach

large (n>30)

$$\sigma^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

Don't often have this much data

Use z tables for this approach

Example

n	Temp
1	40.1
2	39.2
3	43.2
4	47.2
5	38.6
6	40.4
7	37.7

$$\bar{x} = \frac{(40.1 + 39.2 + 43.2 + 47.2 + 38.6 + 40.4 + 37.7)}{7} = 40.9$$

$$s^2 = \frac{1}{7-1}\begin{bmatrix} (40.1-40.9)^2 + (39.2-40.9)^2 + \\ (43.2-40.9)^2 + (47.2-40.9)^2 + \\ (38.6-40.9)^2 + (40.4-40.9)^2 + \\ (37.7-40.9)^2 \end{bmatrix} = 10.7$$

$$s = 3.27$$

Standard Deviation Summary

(normal distribution)

40.9 ± (3.27) 1s: 68.3% of data are within this range

40.9 ± (3.27x2) 2s: 95.4% of data are within this range

40.9 ± (3.27x3) 3s: 99.7% of data are within this range

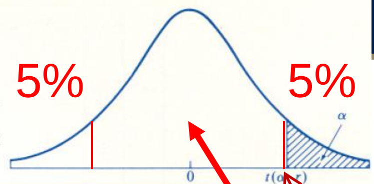
If normal distribution is questionable, use Chebyshev's inequality:

At least 50% of the data are within 1.4 s from the mean.
At least 75% of the data are within 2 s from the mean.
At least 89% of the data are within 3 s from the mean.

Note: The above ranges don't state how accurate the mean is - only the % of data within the given range

Student t-test (gives confidence of where μ (not data) is located)

Table C.6 Upper Percentage Points of the Student's *t*-Distribution: Values of $t(\alpha; r)$



r	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
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
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Source: Reproduced with permission from Table 12 of E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 1 (Cambridge: Cambridge University Press, 1954).

measured mean

$$\mu = \bar{x} \pm t \cdot \left(\frac{s}{\sqrt{n}} \right) \quad \text{where t} = \text{f}\left(\frac{\alpha}{2}, n-1 \right)$$

true mean

α=1- confidence

r = n-1 = 6

2-tail

$$\mu = 40.9 \pm \ ?$$

Remember
s = 3.27

Conf.	α/2	t	+-
90%	.05	1.943	2.40
95%	.025	2.447	3.02
99%	.005	3.707	4.58

t-test in Excel

- The one-tailed t-test function in Excel is:

  =T.INV(α,r)

  - Remember to put in $\alpha/2$ for tests (i.e., 0.025 for 95% confidence interval)

- The two-tailed t-test function in Excel is:

  =T.INV.2T(α,r)

Where

- α is the probability
  - (i.e, .05 for 95% confidence interval for 2-tailed test) and
- r is the value of the degrees of freedom

T-test Summary

μ= exact mean

40.9 is sample mean

40.9 ± 2.4   90% confident μ is somewhere in this range

40.9 ± 3.0   95% confident μ is somewhere in this range

40.9 ± 4.6   99% confident μ is somewhere in this range

Outline

1. Error of Measured Variables
2. Comparing Averages of Measured Variables

Comparing averages of <u>measured</u> variables

Experiments were completed on two separate days.

Day 1:	$\bar{x}_1 = 40.9$	$s_{x1} = 3.27$	$n_{x1} = 7$
Day 2:	$\bar{x}_2 = 37.2$	$s_{x2} = 2.67$	$n_{x2} = 9$

When comparing means at a given confidence level (e.g. 95%), is there a difference between the means?

Comparing averages of measured variables

New formula:

Step 1
(compute *T*)

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_{x1}-1)s_{x1}^2 + (n_{x2}-1)s_{x2}^2}{n_{x1}+n_{x2}-2}\left(\frac{1}{n_{x1}} + \frac{1}{n_{x2}}\right)}}$$

For this example, $T = 2.5$

Larger |*T*|:
More likely
different

Step 2
Compute net *r*

r	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

← $r = n_{x1} + n_{x2} - 2$

Comparing averages of <u>measured</u> variables

Step 3
Compute net *t* from net *r*

r	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

Step 4
Compare |T| with *t*

At a given confidence level (e.g. 95% or α=0.05), there is a difference if:

$$|T| > t\left(\frac{\alpha}{2}, r\right)$$

2-tail

T t

2.5 > 2.145

95% confident there is a difference!

(but not 98% confident)

Example
(Students work in Class)

1. Calculate $\bar{x}$ and s for both sets of data
2. Find range in which 95.4% of the data fall (for each set).
3. Determine range for μ for each set at 95% probability
4. At the 95% confidence level are the pressures different each day?

Data points	Pressure Day 1	Pressure Day 2
1	750	730
2	760	750
3	752	762
4	747	749
5	754	-