

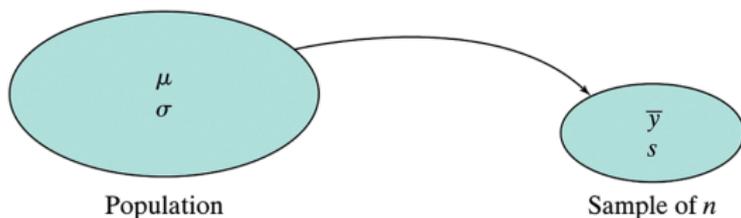
Sections 6.2 and 6.3

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Estimating population parameters



Take a random sample of data Y_1, \dots, Y_n from the population; \bar{y} estimates μ and s estimates σ .

Example 6.1.1 Butterfly wings

$n = 14$ male Monarch butterflies were measured for wing area (Oceano Dunes State Park, California).

Wing area (cm ²)				
33.9	33.0	30.6	36.6	36.5
34.0	36.1	32.0	28.0	32.0
32.2	32.2	32.3	30.0	

$\bar{y} = 32.81$ cm² and $s = 2.48$ cm² estimate μ and σ , the mean and standard deviation of *all Monarch butterfly wing areas from Oceano Dunes*.

How good are these estimates? Can we provide a *plausible range* for μ ?

6.2 Standard error of \bar{Y}

- Recall on p. 151 that $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$.
- We will usually not know σ (if we don't know μ , how can we know σ ?)
- Simply plug in s for σ .
- The **standard error of the mean** is

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

- For the butterfly wings, $SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{2.48}{\sqrt{14}} = 0.66 \text{ cm}^2$.
- The standard error $SE_{\bar{Y}}$ gives the variability of \bar{Y} ; the standard deviation s gives the variability *in the data itself*.

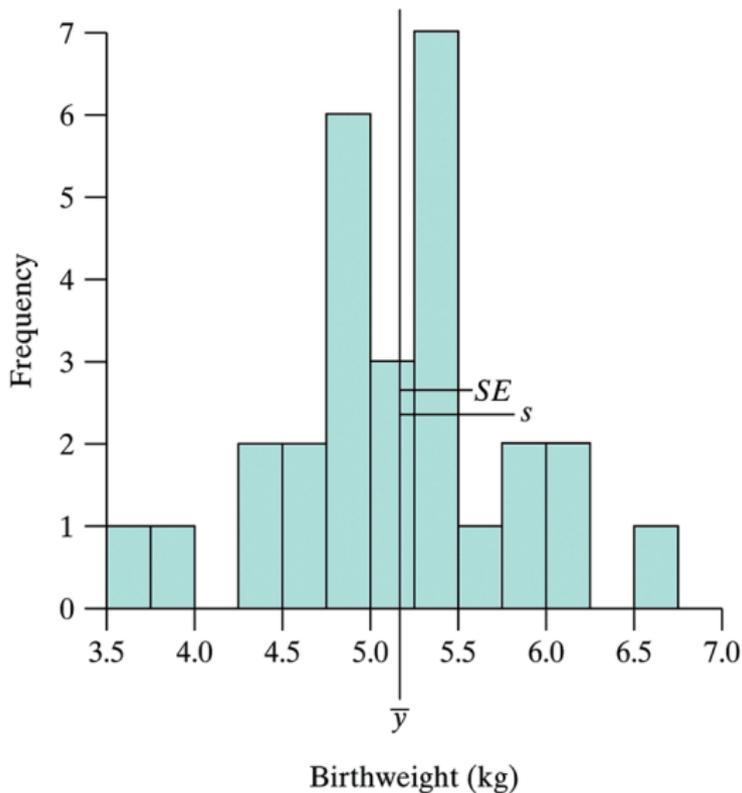
Example 6.2.2

Geneticist weighs $n = 28$ female Rambouillet lambs at birth, all born in April, all single births.

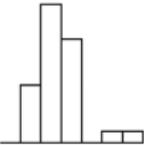
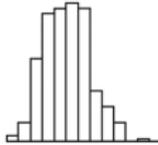
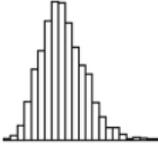
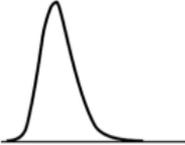
Birthweight (kg)						
4.3	5.2	6.2	6.7	5.3	4.9	4.7
5.5	5.3	4.0	4.9	5.2	4.9	5.3
5.4	5.5	3.6	5.8	5.6	5.0	5.2
5.8	6.1	4.9	4.5	4.8	5.4	4.7

- $\bar{y} = 5.17$ kg estimates μ , the population mean.
- $s = 0.65$ kg estimates the spread *in the sample*.
- $SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0.65}{\sqrt{28}} = 0.12$ kg estimates how variable \bar{y} is, i.e. how “close” we can expect \bar{y} to be to μ .

Birthweight of $n = 28$ lambs

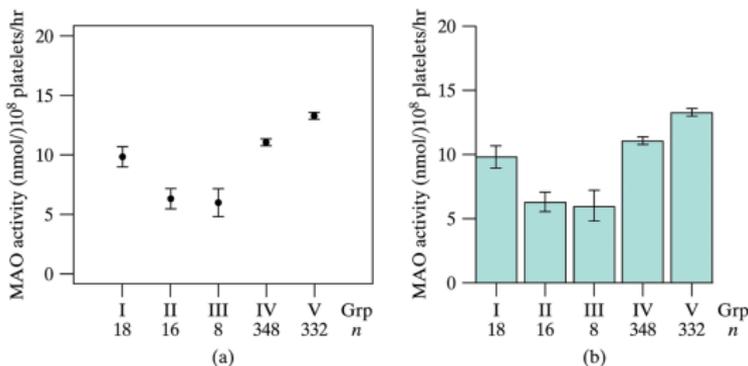


Increasing n sampling from lamb birthweight population

	$n = 28$	$n = 280$	$n = 2,800$	$n \rightarrow \infty$
\bar{y}	5.17	5.19	5.14	$\bar{y} \rightarrow \mu$
s	0.65	0.67	0.65	$s \rightarrow \sigma$
SE	0.12	0.040	0.012	$SE \rightarrow 0$
Sample distribution				

Example 6.2.4 MOA data using SE 's across groups

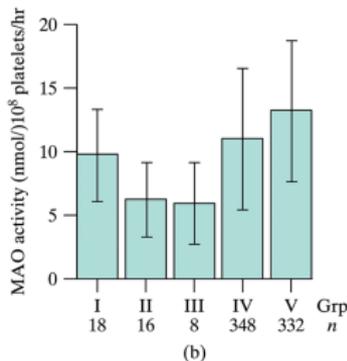
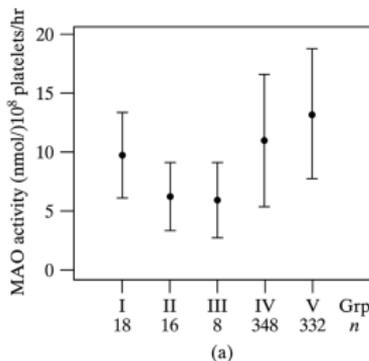
MOA levels vs. schizophrenia diagnosis (I, II, III) and healthy male and female controls (IV and V).



$\bar{y} \pm SE$ using (a) an interval plot, and (b) a bargraph with standard error bars. Gets at how variable the sample means are.

Example 6.2.4 MOA data using s 's across groups

MOA levels vs. schizophrenia diagnosis (I, II, III) and healthy male and female controls (IV and V).



$\bar{y} \pm s$ using (a) an interval plot, and (b) a bargraph with standard deviation bars. Gets at how variable the data are.

Example 6.2.4 MOA data table with all information

MAO activity (nmol/10 ⁸ platelets/hr)				
Group	<i>n</i>	Mean	SE	SD
I	18	9.81	0.85	3.62
II	16	6.28	0.72	2.88
III	8	5.97	1.13	3.19
IV	348	11.04	0.30	5.59
V	332	13.29	0.30	5.50

Confidence interval in one minute...

- \bar{y} provides an estimate of μ , but often we'd like a plausible range for μ .
- Theorem 5.2.1 (p. 151) tells us \bar{Y} is $N(\mu, \frac{\sigma}{\sqrt{n}})$. This holds perfectly when the data Y_1, \dots, Y_n are normal, otherwise it's approximate.
- We can estimate $\frac{\sigma}{\sqrt{n}}$ by $SE_{\bar{Y}}$.
- The 68/95/99.7 rule says that any normal random variable is within 2 standard deviations of its mean 95% of the time.
- **Therefore** \bar{Y} is within $2SE_{\bar{Y}}$ of μ 95% of the time.
- **Restated** μ is within $2SE_{\bar{Y}}$ of \bar{Y} 95% of the time.
- A quick, rough confidence interval for μ is $(\bar{Y} - 2 SE_{\bar{Y}}, \bar{Y} + 2 SE_{\bar{Y}})$.

Confidence interval, known σ , formal derivation

Say we know σ (for now) and the data are normal. Then

$$\bar{Y} \sim N(\mu, \sigma_{\bar{Y}}) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

We can standardize \bar{Y} to get

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}.$$

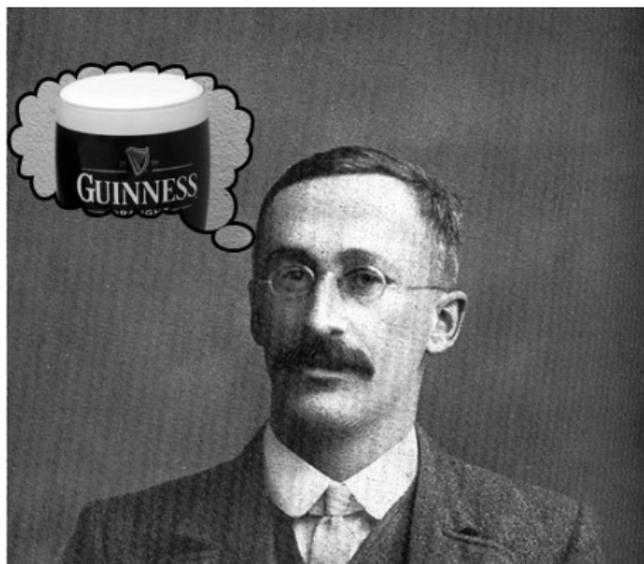
We can show $\Pr\{-1.96 \leq Z \leq 1.96\} = 0.95$. Then

$$\begin{aligned} 0.95 &= \Pr\{-1.96 \leq Z \leq 1.96\} \\ &= \Pr\left\{-1.96 \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right\} \\ &= \Pr\left\{-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{Y} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right\} \\ &= \Pr\left\{\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}\right\} \end{aligned}$$

Confidence interval

- $\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ is a 95% probability interval for μ .
- Once we go out and see $\bar{Y} = \bar{y}$, e.g. $\bar{y} = 32.8 \text{ cm}^2$, there is no probability. Either the interval includes μ or not (more in a minute...)
- We don't actually know $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$, but we do know $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$.
- William Sealy Gosset figured out what $\frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$ is distributed as.

William Sealy Gosset, brewer & statistician



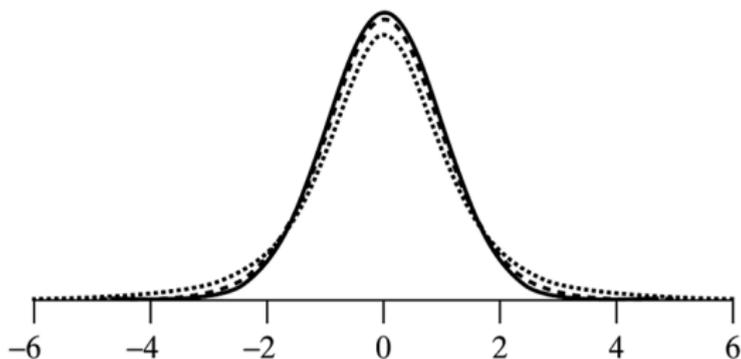
“Mmmmmmmmm...Guinness.”

The t distribution was published by Gosset in 1908 & related to quality control at Guinness brewery.

Estimating σ by s gives a t distribution

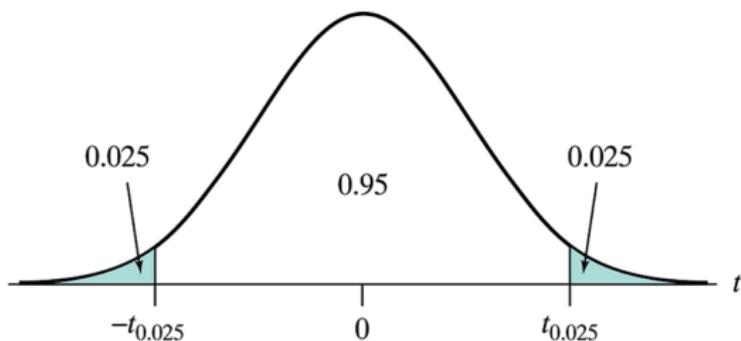
- Instead of normal, $\frac{\bar{Y}-\mu}{SE_{\bar{Y}}}$ has a **Student's t distribution** with $n - 1$ degrees of freedom.
- The student's t distribution looks like a standard normal, but has fatter tails to account for extra variability in estimating $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ by $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$.
- However, the confidence interval is computed the same 'formal' way, replacing $\sigma_{\bar{Y}}$ by $SE_{\bar{Y}}$ and using a t distribution rather than a normal.
- R takes care of the details for us! `t.test(data)` gives a 95% CI for μ .
- For small sample sizes ($n < 30$, say), data need to be approximately normal, otherwise the central limit theorem kicks in.

Two student's t curves (df=3 & 10), and normal curve



t distributions have slightly fatter tails to account for estimating σ by s .

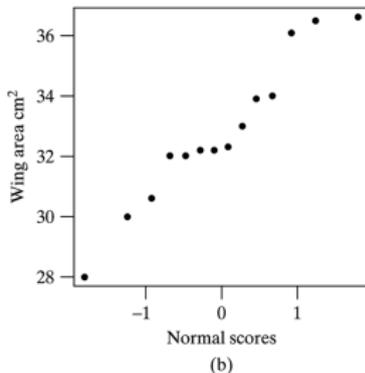
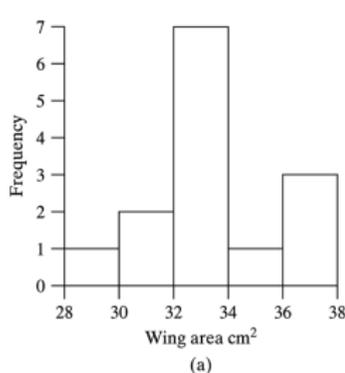
Definition of critical value $t_{0.025}$



We replace “1.96” (from a normal) by the equivalent t distribution value, denoted $t_{0.025}$. Table of these on back inside cover.

Example 6.3.1 butterfly data

Wing area of $n = 14$ male Monarch butterfly wings at Oceano Dunes in California.



This is a small sample size ($n < 30$). We need to check if the data are normal to trust the confidence interval; the histogram looks roughly bell-shaped and the normal probability plot looks reasonably straight.

Confidence interval in R using `t.test`

```
> butterfly=c(33.9,33.0,30.6,36.6,36.5,34.0,36.1,32.0,28.0,32.0,32.2,32.3,32.3,30.0)
> par(mfrow=c(1,2))
> hist(butterfly)
> qqnorm(butterfly)
> t.test(butterfly)
```

One Sample t-test

```
data: butterfly
t = 49.6405, df = 13, p-value = 3.292e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 31.39303 34.24983
sample estimates:
mean of x
 32.82143
```

The part we care about right now is just

```
95 percent confidence interval:
 31.39303 34.24983
```

We are 95% confident that the true population mean wing area is between 31.4 and 34.2 cm^2 .

Other confidence levels

- Sometimes people want a 90% CI or a 99% CI. As confidence goes up, the interval *must become wider*. To be *more confident* that the mean is in the interval, we need to include more plausible values.
- The corresponding multipliers are $t_{0.05}$, $t_{0.025}$, and $t_{0.005}$ for 90%, 95%, and 99% CI's, respectively. These are in the table on the inside cover of the back of your book if you construct a CI by hand.
- In R, use `t.test(data,conf.level=0.90)` for a 90% test CI
`t.test(data,conf.level=0.99)` for 99% CI.

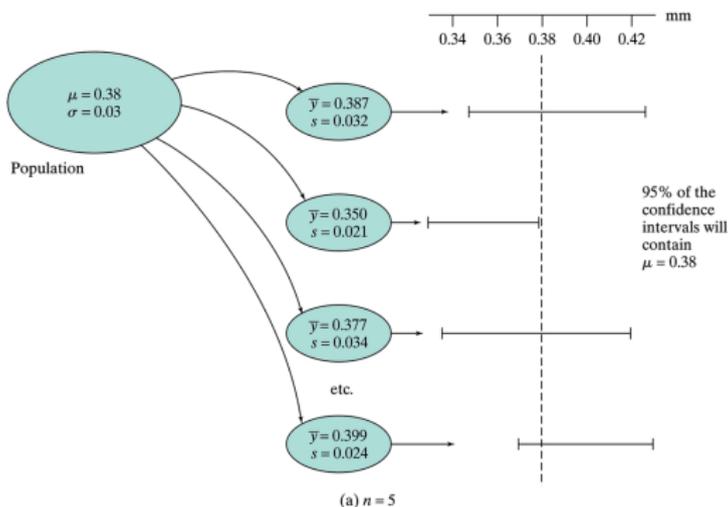
```
> t.test(butterfly,conf.level=0.9)
90 percent confidence interval:
 31.65052 33.99234
> t.test(butterfly)
95 percent confidence interval:
 31.39303 34.24983
> t.test(butterfly,conf.level=0.99)
99 percent confidence interval:
 30.82976 34.81309
```

Interpretation of CI

- The CI $\bar{Y} \pm t_{0.025}SE_{\bar{Y}}$ is *random* until we see $\bar{Y} = \bar{y}$.
- Then the CI either covers μ or not, *and we don't know which!*
- After we compute the observed CI, we talk about “confidence” not “probability” (bottom, p. 181).
- If we did a meta-experiment and collected samples of size n repeatedly and formed 95% CI's, approximately 95 in 100 would cover μ .
- Increasing n only makes the intervals smaller; still 95% of the CI's would cover μ .
- *However, we only get to see one of these intervals, because we only take one sample.*

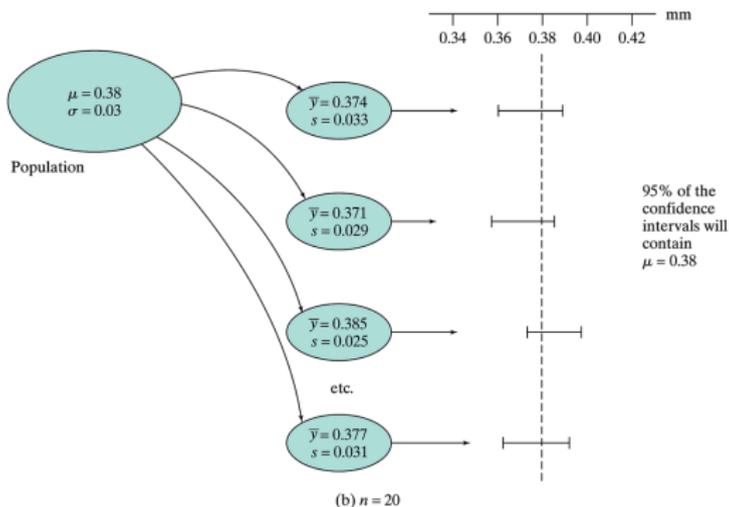
Eggshell thickness $n = 5$

Meta-experiment for eggshell thickness where $\mu = 0.38$ mm & $\sigma = 0.03$ mm.



Eggshell thickness $n = 20$

Meta-experiment for eggshell thickness where $\mu = 0.38$ mm & $\sigma = 0.03$ mm.



Invisible man walking his dog

A confidence interval is like an invisible man walking his dog.

We can see the dog (\bar{y}) one time (one sample) and know that the dog is within two standard errors of the mean $2SE_{\bar{y}}$ of the invisible man μ with 95% probability at any given time. So we're pretty confident that the invisible man is within $2SE_{\bar{y}}$ of the dog.

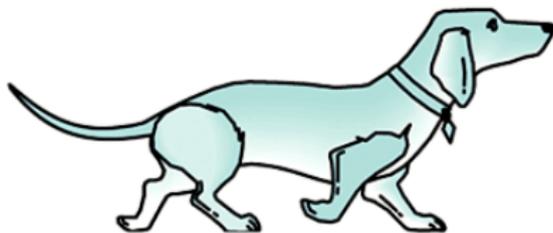


Figure 6.3.1 Invisible man walking his dog

Review

- A *confidence interval* provides a plausible range for μ .
- Since \bar{Y} is normal, the 68/95/99.7 rule says μ is within $\bar{Y} \pm 2SE_{\bar{Y}}$ 95% of the time.
- This interval is too small; Gosset introduced the t distribution to make the interval more accurate $\bar{Y} \pm t_{0.025}SE_{\bar{Y}}$; `t.test(sample)` in R takes care of the details.
- For $n < 30$ the data must be normal; check this with normal probability plot. For $n \geq 30$ don't worry about it.
- Interpretation is important. "With 95% confidence the true mean of [population characteristic] is between [a] and [b] [units]."