Sections 6.2 and 6.3

Note made by: Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Review

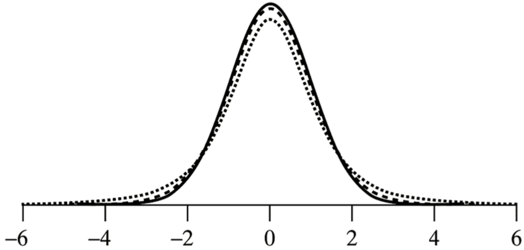
- $\overline{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is a 95% probability (? or confidence?) interval for μ. Or, $(\overline{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{Y} + 1.96\frac{\sigma}{\sqrt{n}})$.
- By empirical rule, a quick, rough confidence interval for μ is $(\overline{Y} - 2\frac{\sigma}{\sqrt{n}}, \overline{Y} + 2\frac{\sigma}{\sqrt{n}})$.
- If σ is known and data normal, then

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

- What if σ is unknown?
- What if data is non-normal?
- Recall: $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ and $\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$

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 students 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 $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$.
- Two student's t curves (df$=$ 3&10) vs. Standard normal curve.

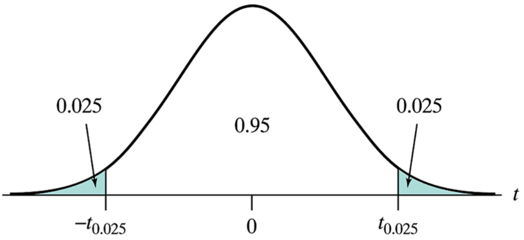


Estimating σ by s gives a t distribution

- However, the confidence interval is computed the same way, replacing $\sigma_{\bar{Y}}$ by $SE_{\bar{Y}}$ and using a t distribution rather than a normal.
- For small sample sizes ($n < 30$, say), data need to be approximately normal, otherwise the central limit theorem kicks in.

Definition of critical value $t_{0.025}$

t_α is defined so that $\Pr\{T > t_\alpha\} = \alpha$ where $T \sim t_{df}$.



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

95 % confidence interval for μ, when σ is unknown.

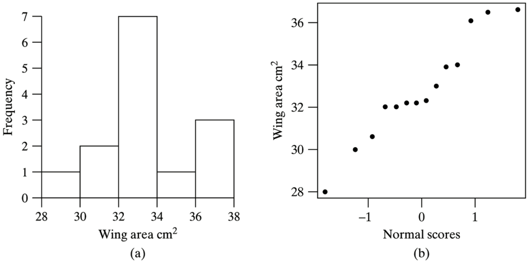
Summary: we estimate unknown σ with S, and use a t distribution rather than standard normal, the 95% CI is given by

$$\overline{Y} \pm t_{0.025} \frac{S}{\sqrt{n}},$$

R takes care of the details for us! t.test(data) gives a 95% CI for μ.

Example 6.3.1 butterfly data

Wing area of $n = 14$ male Monarch butterly 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.

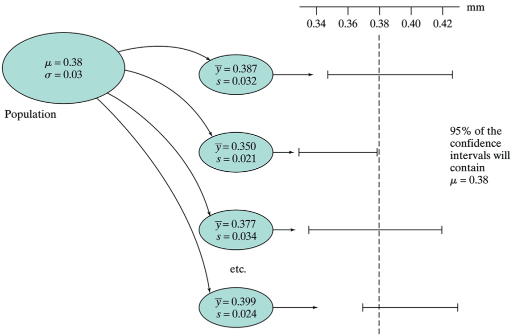
```
> t.test(butterfly,conf.level=0.9)
90 percent confidence interval:
 31.65052 33.99234
> t.test(butterfly)
95 percent confidence interval:
 31.39303 34.24983
> t.test(butterfly,conf.level=0.99)
99 percent confidence interval:
 30.82976 34.81309
```

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 <u>one</u> of these intervals, because we only take one sample.*

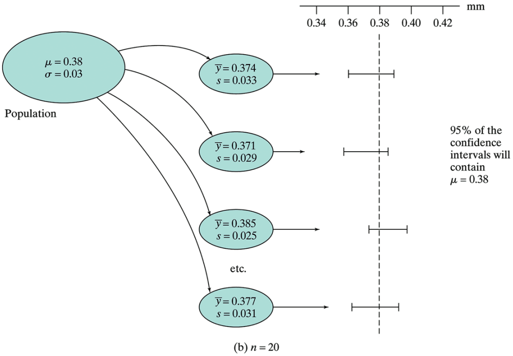
Eggshell thickness $n = 5$

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



(a) $n = 5$

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 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.

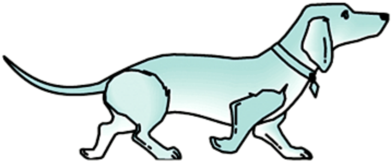


Figure 6.3.1 Invisible man walking his dog

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 $\boxed{\text{population characterstic}}$ is between $\boxed{\text{a}}$ and $\boxed{\text{b}}$ $\boxed{\text{units}}$."