Ch. 1 - sample ⟷ Ch. 3-7 theory.

$\bar{x}$, s^2 μ, σ^2

Chapter 8 Statistical Modelling and Inference

Statistical Inference: the process of inferring something about the population based on what is measured in the sample.

We will learn:

$\bar{x}$, s^2 μ, σ^2

1. **Point Estimation** - estimate an unknown parameter using a single number calculated from the sample data
2. **Confidence Intervals** - estimate an unknown parameter using an interval of values that is likely to contain the true value of that parameter
3. **Hypothesis Testing** - we have some claim about the population, and we check whether or not the sample data provide evidence against this claim

Point Estimates for μ and σ

A statistic used to estimate a parameter is an **unbiased estimator** of the parameter if the mean of the sampling distribution is equal to the true value of the parameter. $\hat{\theta}$ is an unbiased estimator of θ if:

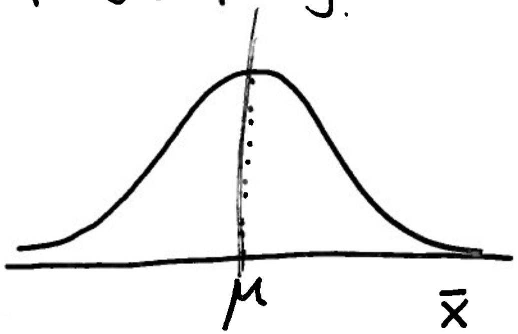
$$E(\hat{\theta}) = \theta \leftarrow \text{parameter.}$$

Suppose that $X_1, ..., X_n$ is a simple random sample from a population with mean $\underline{\underline{\mu}}$ and variance $\underline{\underline{\sigma^2}}$.

- $\bar{x}$ is an unbiased estimator of μ: $E(\bar{X}) = \mu$

$$E(\bar{X}) = E(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n}\sum_{i=1}^{n} \underline{\underline{E(X_i)}} = \frac{1}{\cancel{n}}\cancel{n} \times \mu = \underline{\underline{\mu}}$$

$\bar{X}$ sampling distribution.



long term avg of our estimate is exactly what we want to estimate

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

↖

Real World

data
(sample)

$\Downarrow$

statistic

Theoretical World

Statistical + Scientific models
(population)

$\Downarrow$

parameter

statistic $\xrightarrow{\text{estimate}}$ parameter

eg. IQ levels of students at UBC

μ mean IQ level of all UBC students. (unknown)

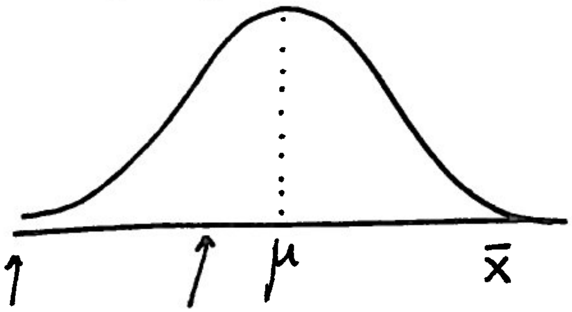
$\bar{x}$ point estimate, sample mean IQ of sample of 100 students.

Ch. 7

sampling · dist of $\bar{x}$

$E(\bar{x})$

$\bar{x}$



- $\underline{\underline{s^2}}$ is an unbiased estimator of $\underline{\underline{\sigma^2}}$:

$$E(s^2) = \sigma^2$$

Notice we can rewrite s^2:

$$\begin{aligned}
s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1 \leftarrow} \qquad \text{Ch. 1} \\
&= \frac{\sum x_i^2 - 2\boxed{\sum x_i}\,\bar{x} + n\bar{x}^2}{n-1} \\
&= \frac{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}{n-1} \\
&= \frac{\sum x_i^2 - n\bar{x}^2}{n-1}
\end{aligned}$$

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

$$n\bar{x} = \sum x_i$$

Proof:

$$E(S^2) = E\left[\frac{\sum X_i^2 - n\bar{X}^2}{n-1}\right]$$

$$= \frac{1}{n-1}\left[n\underbrace{E[X_i^2]} - n\underbrace{E[\bar{X}^2]}\right]^*$$

$$= \frac{n}{n-1}\left[\sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{n} - \cancel{\mu^2}\right]$$

$$= \frac{\cancel{n}}{n-1}\left[\frac{n\sigma^2 - \sigma^2}{\cancel{n}}\right]$$

$$= \frac{\sigma^2\cancel{(n-1)}}{\cancel{n-1}} = \sigma^2$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1 \leftarrow}$$

Recall:

$$\mathrm{Var}(X) = E(X^2) - E(X)^2$$

$$\sigma^2 = E(X^2) - \mu^2$$

$$E(X^2) = \sigma^2 + \mu^2$$

$$\mathrm{Var}(\bar{X}) = E[\bar{X}^2] - E[\bar{X}]^2$$

$$\frac{\sigma^2}{n} = E[\bar{X}^2]^* - \mu^2$$

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

Recall:

Sampling distribution of the sample mean

Suppose x is a variable with a certain distribution with population mean μ and standard deviation σ.

- If n is large enough and other conditions are satisfied, then the sampling distribution of the sample means

$$\bar{X} \overset{\text{approx}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

  (Ch. 7)
- If the underlying distribution of x is Normal, the sampling distribution of $\bar{X}$ is Normal and n need not be large. (Ch. 5)

Confidence interval for the population mean μ

A **confidence interval** is used to describe the amount of uncertainty associated with a sample estimate of a population parameter. It consists of a range of values that act as plausible estimates of the population parameter. In general, a **confidence interval** takes the form:

$$\text{Estimate} \pm \text{Margin of Error}$$

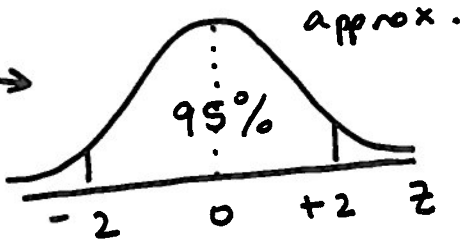
The uncertainty associated with the confidence interval is specified by the confidence level (C). Common confidence levels are 90%, 95% and 99%.

$E(\bar{X}) = \mu$

$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$\underset{\uparrow}{\bar{x}} \pm$ margin error.

Using 68-95-99.7 rule, →



$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| > 2\right) \simeq 0.05$$

$$P\left(-2 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 2\right) \simeq 0.95$$

$$P\left(\bar{X} - \underbrace{2\frac{\sigma}{\sqrt{n}}}_{} \le \underset{\substack{\uparrow \\ \text{fixed.}}}{\mu} \le \bar{X} + \underbrace{\underset{\substack{\uparrow \\ z^*}}{2}\frac{\sigma}{\sqrt{n}}}_{}\right) \simeq \underline{\underline{0.95}}$$

← $\underline{\underline{95\%}}$ confident population mean μ falls within $\frac{2\sigma}{\sqrt{n}}$ units of $\bar{x}$

$\underline{\underline{\bar{X}}}$ random

One-sample z-interval

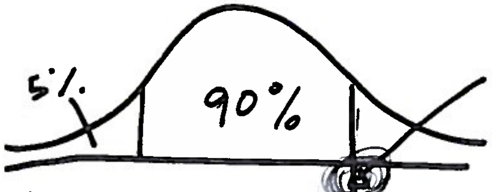
- When σ is known we use $SD(\bar{x})$ and the normal model to construct confidence intervals for μ. A confidence interval of confidence level C for μ is computed using:

$$\boxed{\bar{x} \pm \underbrace{z^* \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}}$$

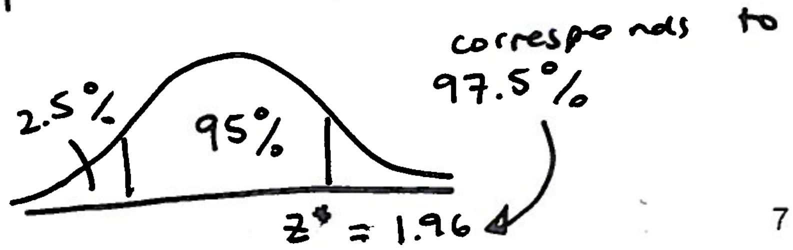
critical value

$SD(\bar{X})$

point estimator



z^*	1.645	1.96	2.576
C	90%	95%	99%



Example 1

Polychlorinated biphenyls (PCBs) are industrial chemicals synthesized in North America in 1929. They were used in manufacturing of electrical equipment, hydraulic systems, and other applications. PCBs were banned in Canada in the 1970s because of environmental concerns. Despite the reductions in PCB, releases of PCBs to the environment through spills and fires continue to occur.

Suppose samples are taken from a lake and the amount of PCB (in μg/l) is Normally distributed with true standard deviation 0.75 μg/l.

(a) Compute the 95% confidence interval for the true mean PCB content if the average PCB content for 20 water samples was 4.85 μg/l.

(b) Suppose the researchers would like to conduct a second study, but with margin of error 0.2 and 99% confidence. How many lake samples would the researchers need to take?

a) 95% CI. X = amount of PCB $\sim N(\mu, \underline{\underline{\sigma = 0.75}})$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{x} = 4.85$ μg/l

$n = 20$

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.75}{\sqrt{20}}$$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 4.85 \pm 1.96 \times \frac{0.75}{\sqrt{20}}$$

$$= 4.85 \pm \underbrace{0.33}_{ME} = (4.52, 5.18)$$

We are 95% confident true mean PCB content is between 4.52 and 5.18 μg/l

b) n = ?

$$ME = z^* \frac{\sigma}{\sqrt{n}}$$

$$0.2 = 2.575 \times \frac{0.75}{\sqrt{n}}$$

$$n = \left(\frac{2.575 \times 0.75}{0.2}\right)^2 = 93.24$$

→ 94 (round up)

Last Class:

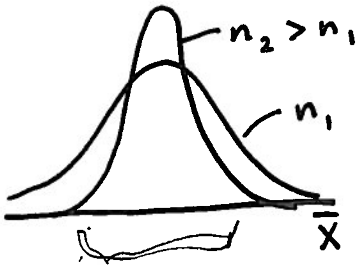
Confidence intervals: $\boxed{\text{estimate} \pm \text{margin of error}}$

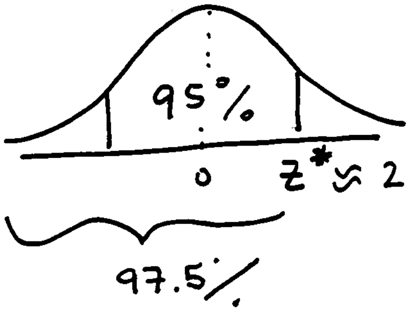
one sample z-interval: $\bar{x} \pm \overbrace{\underbrace{z^*}_{\text{critical value}} \underbrace{\frac{\sigma}{\sqrt{n}}}_{SD(\bar{X})}}^{ME}$

n large

$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}} = SD(\bar{X})\right)$

by CLT (Ch. 7)





What if σ is unknown? In most situations, the population standard deviation σ is not known.

$\underbrace{\bar{x}}_{\text{random}}, \underbrace{\mu}_{\text{fixed}} \qquad \underbrace{s}_{\text{random}} \quad \underbrace{\sigma}_{\text{fixed.}}$

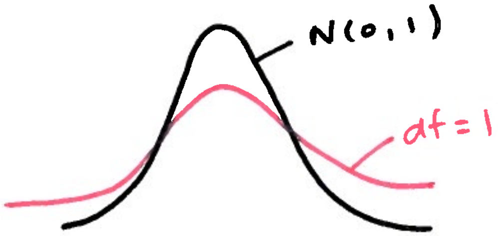
Standard error:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

sample sd.

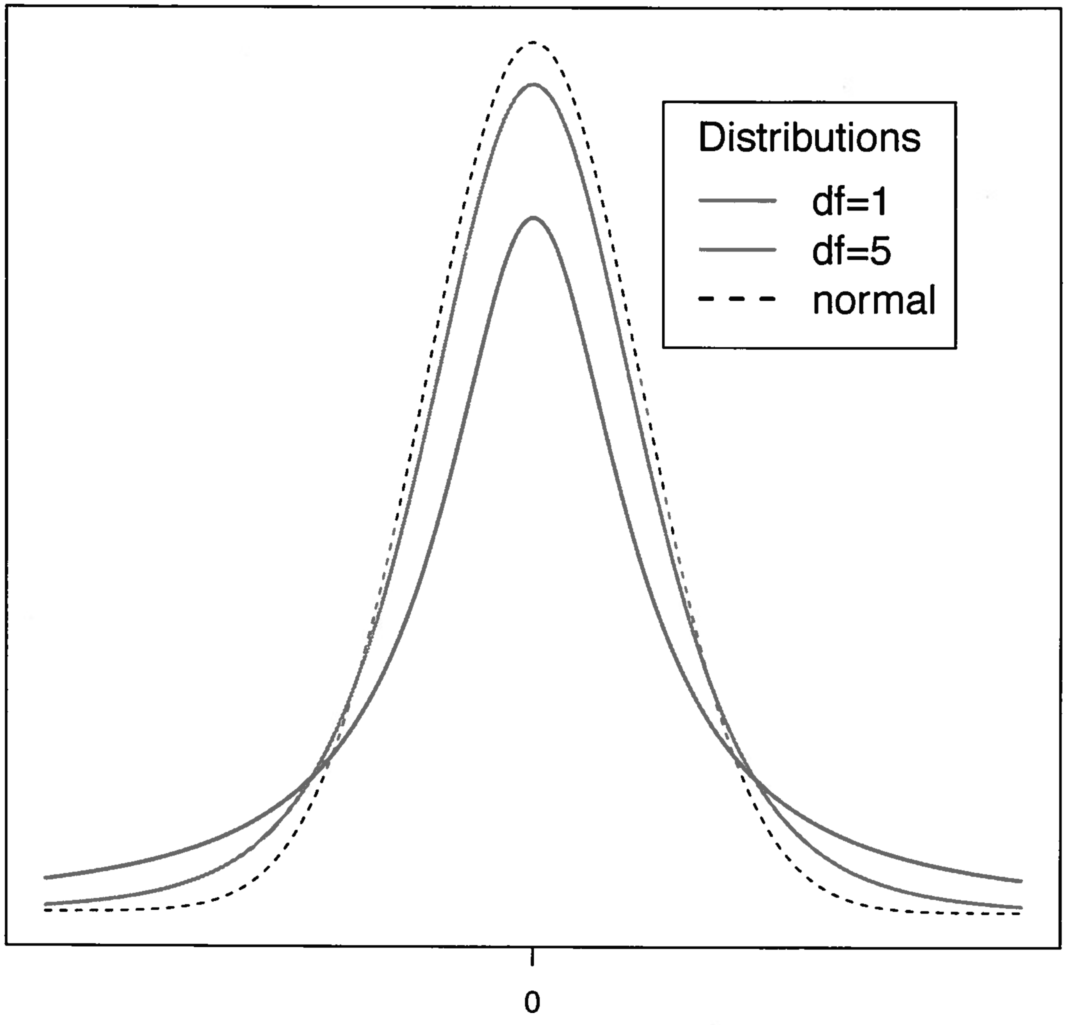
to estimate the variability of $\bar{x}$.

- For a particular $\bar{x}$ value the corresponding $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1)$.
- If σ is unknown and $SE(\bar{x}) = \frac{s}{\sqrt{n}}$ is used to estimate $\frac{\sigma}{\sqrt{n}}$ then the quantity $\boxed{\frac{\bar{x}-\mu}{SE(\bar{x})}}$ may not be well described by the standard Normal model.
- Thus we introduce a distribution, similar to the Normal distribution, which now accounts for the sample size. The sampling distribution of t is called the **Student's t-model**.



$$\boxed{t = \frac{\bar{x} - \mu}{s/\sqrt{n}}} \quad SE(\bar{X})$$

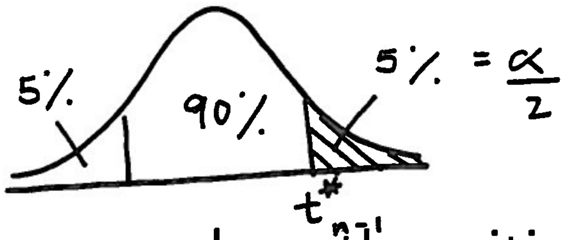
The **Student's t-model** is unimodal, bell-shaped and symmetric about the mean 0. There is one model parameter called **degrees of freedom** $(df = n - 1)$, which determines the shape of the model.



One-sample t-interval

- When σ is unknown, we use $SE(\bar{x})$ and the t-model to construct confidence intervals for μ. A confidence interval of confidence level $\underline{\underline{C}}$ for μ is computed using:

$\alpha = 100 - C$
$= 10\%$



$$\boxed{\bar{x} \pm \underbrace{\overbrace{t^*_{\underline{\underline{n-1}}}}^{\text{critical value}} \overbrace{\frac{s}{\sqrt{n}}}^{SE(\bar{x})}}}$$

estimate ± ME

sample sd

where the critical value t^*_{n-1} is the t-value with an area of $(100 - C)$ in the tails and $n - 1$ degrees of freedom

Note: For large samples, $\underline{t_{n-1}}$ and $N\underline{(0, 1)}$ are almost identical and we can use $\underline{\underline{z}}$ as the test statistic.
For <u>small samples</u>, $\underline{t_{n-1}}$ and $\underline{N(0, 1)}$ are different and we use $\underline{\underline{t}}$ as the test statistic.

n large 30.

Example 2

Suppose a machine in a Coca Cola factory is used to fill containers with Coke. Factory managers want to montitor the amount that is being dispensed. A random sample of 10 containers is selected and the contents are shown below. The histogram of the sample is unimodal and roughly symmetric.

25.5 oz	26.1 oz
26.8	23.2
24.2	28.4
25.0	27.8
27.3	25.7

(a) ~~By looking at the displays below, is it appropriate to analyze these data using methods based on Normal distributions?~~

(b) Find the margin of error for 90% confidence.

(c) Construct a 90% confidence interval on the mean fill volume.

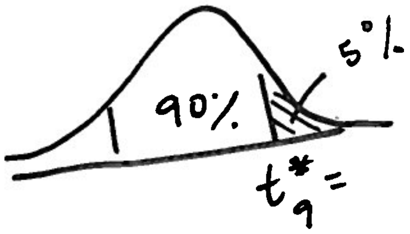
b) $\bar{X} = \dfrac{\sum x_i}{n} = \dfrac{25.5 + \ldots + 25.7.}{10} = 26$

$s^2 = \dfrac{1}{n-1} \sum (x_i - \bar{x})^2 = \dfrac{1}{9}\left[(25.5-26)^2 + \ldots + (25.7-26)^2\right]$

$= 2.64$

- hist. unimodal + symmetric.
- $n = 10$ sample small.
- σ unknown

] t-interval.



$ME = t^*_{n-1} \, SE(\bar{x}) = \underline{\underline{t^*_9}} \dfrac{s}{\sqrt{n}} = \dfrac{1.833\sqrt{2.64}}{\sqrt{10}} = 0.942$

c) 90% CI.

$\bar{x} \pm ME = 26 \pm 0.942 = (\underline{\underline{25.06, 26}}.942)$ ↙

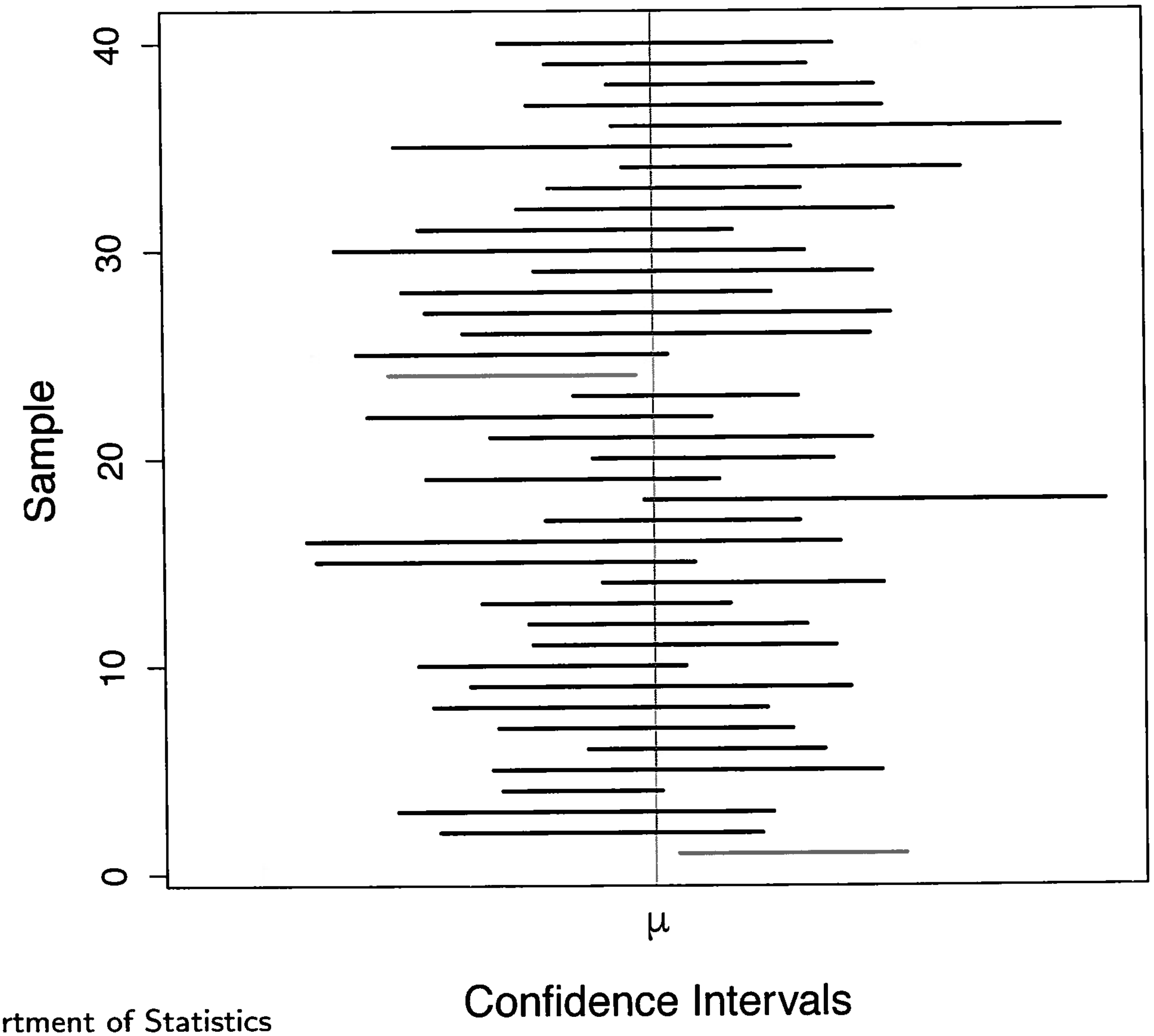
We are <u>90% confident</u> the true mean fill volume is between 25.06 oz and 26.942 oz.

Conditions for constructing confidence intervals:

- the sample is randomly drawn from the population
- the sampled values are independent
- nearly Normal condition
    - if the underlying distribution of x is (nearly) Normal, or unimodal and symmetric then use of the t-model is okay if n is small
    - if the underlying distribution is non-Normal or unknown, we need a large n

Interpreting a confidence interval for μ

If repeated samples were taken and the 95% confidence interval is computed for each sample, 95% of the intervals would contain the population mean.



Testing of Hypotheses about μ

In statistics, a **hypothesis** is a statement or claim about a parameter. We test this claim with a hypothesis test.

Motivating Example

A medium size latte at a certain coffee chain is supposed to be 16 oz. Customers have complained that the company is purposely underfilling lattes. The amount in each latte varies slightly from cup to cup. To determine whether there is sufficient evidence to support the suspicion, we could take a random sample of, say, 100 cups. We could conduct a hypothesis test about our true population mean using the average amount from our sample. Suppose our sample has an average amount of 15.8oz. Does this difference indicate the company is shortchanging customers?

Step 1: Null and alternative hypotheses

- The **null hypothesis** (H_0) is a statement about the value of the population parameter

  $$H_0 : \text{population parameter } = \text{some hypothesized value}$$

  The null hypothesis asserts that an observed difference is due to chance variation. A hypothesis test always begins by assuming the null hypothesis is correct. Then we evaluate the evidence against the null hypothesis.

- The **alternative hypothesis** (H_A) is a statement that opposes the null hypothesis. It asserts that an observed difference is real.

  Suppose we want to test hypotheses about our population mean μ, we have 3 different tests we could perform:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu \neq \mu_0 \qquad \text{(two-tailed)}$$
$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu > \mu_0 \qquad \text{(right-tailed)}$$
$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu < \mu_0 \qquad \text{(left-tailed)}$$

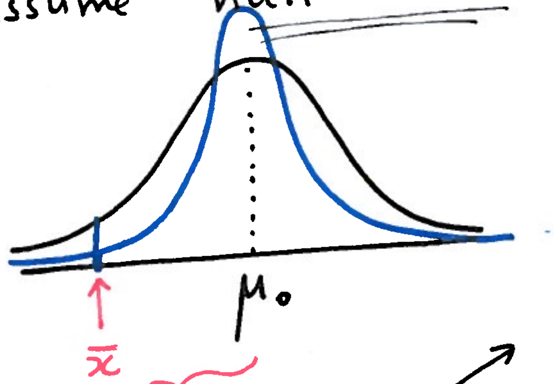
Last Class:

Hypothesis testing:

1) Hypotheses: $H_0: \underline{\underline{\mu = \mu_0}}$ vs. $H_A: \mu \neq \mu_0$] two tailed test

$H_A: \mu > \mu_0$
$H_A: \mu < \mu_0$ } one tailed tests

2) Assume null model,



$\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$ n large, x Normal

difference → due chance?
→ real effect / significant effect?

test statistic: $t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ or $z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

large test statistic → Reject null hypothesis.

Small test stat → not enough evidence against null
do not reject Ho.

Step 2: Test statistic

- Summarize the data into a **test statistic**. A test statistic is constructed assuming the null model is correct.

Case 1: σ is unknown

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

(This test is called the one-sample t-test)

Case 2: σ is known

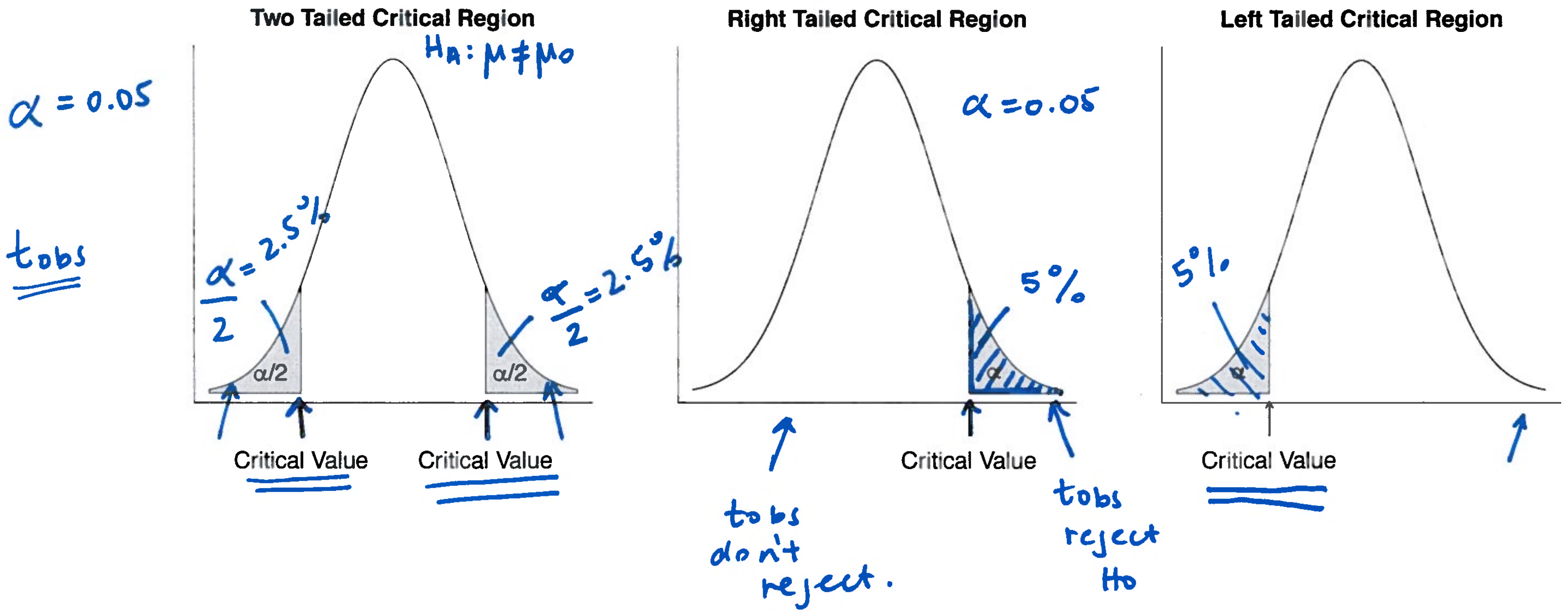
$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

(This test is called the one-sample z-test)

Step 3: Critical region

$\alpha = 0.1$ or $\alpha = 0.01$

- The **significance level** of a test is denoted by α (e.g. $\alpha = 0.05$) establishes a cut-off for making a decision about the null hypothesis.
- The **critical region** consists of outcomes that are very unlikely to occur if the null hypothesis is true.



Step 4: Conclusion

- If the test statistic is large enough to be in the critical region, we reject the null hypothesis.
- If the test statistic has a low value and does not lie in the critical region, we conclude that the evidence from the sample is not sufficient, and the decision is fail to reject the null hypothesis.

Problem 8.2 from course text

Example 3

The time for a worker to repair an electrical instrument is a normally distributed $N(\mu, \sigma^2)$ random variable measured in hours, where both μ and σ^2 are unknown. The repair times for 10 such instruments chosen at random are as follows:

$$212, 234, 222, 140, 280, 260, 180, 168, 330, 250$$

(a) Calculate the sample mean and the sample variance of the 10 observations.

(b) Construct a 95% confidence interval for μ.

(c) Suppose the worker claims that his average repair time for the instrument is no more than 200 hours. Test if there is sufficient evidence to dispute the worker's claim.

a)
$$\left.\begin{aligned} \bar{x} &= 212 + \dots + 250 / 10 = 227.60 \\ s^2 &= \frac{(212 - 227.60)^2 + \dots + (250 - 227.60)^2}{9} = 3176.71 \end{aligned}\right] \text{Ch. 1}$$

b) Thought process: $\left.\begin{array}{l} \sigma^2 \text{ unknown} \\ \text{small sample size} \end{array}\right]$ use $\underline{\underline{t}}$

→ told repair times Normal.

→ Random, independent.

95% CI: estimate ± ME

$$\bar{x} \pm t^*_{n-1} \; SE(\bar{x})$$

$$\bar{x} \pm t^*_9 \frac{s}{\sqrt{n}} = \underline{\underline{227.60}} \pm 2.262 \frac{\sqrt{3176.71}}{\sqrt{10}}$$

$$= (187.28, 267.92) \text{ hours.}$$

Were 95% confident the true mean repair time is between 187.28 and 267.92 hours.

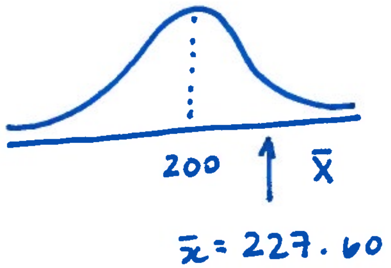
c) ① $H_0: \mu = 200$ mean repair time for all instruments is 200 hours

$H_A: \mu > 200$ mean repair time for all instruments is greater than 200 hours.

$\uparrow$

worker claims his avg. repair time is ≤ 200
but we want to dispute.

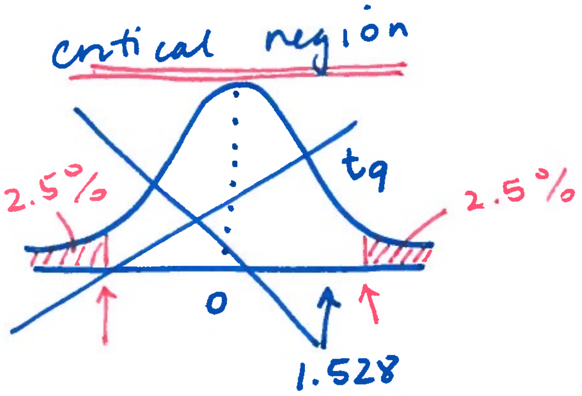
② Test stat: Under null model,

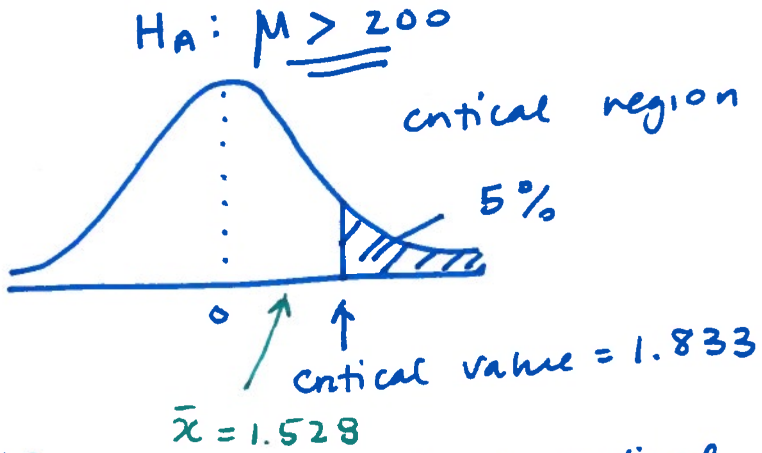


$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{227.60 - 200}{\sqrt{3176.71}/\sqrt{10}}$$

$$= 1.528$$

③ $\alpha = \underline{\underline{0.05}}$





④ $t_{obs} = 1.528$ is less than 1.833 and is not in critical region. We do not reject H_0.
There is not enough evidence to suggest the true mean repair

is significantly greater than 200 hours at 5% significance level.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{264 - 269}{10/\sqrt{100}} = -5.$$

Two-sided significance tests and confidence intervals

$H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \underline{\underline{\mu_0}}$

In a two-sided hypothesis test of the population mean at level α, the null hypothesis $H_0 : \mu = \mu_0$ (against $H_A : \mu \neq \mu_0$) is:

- rejected if the $100(1-\alpha)\%$ confidence interval for μ <u>does not include</u> the hypothesized mean (μ_0).
- not rejected if the $100(1-\alpha)\%$ confidence interval for μ <u>does include</u> $\underline{\underline{\mu_0}}$.

$H_0 : \mu = 15$

Steps:

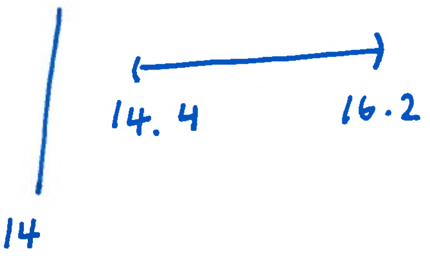
1. Set up null and alternative hypotheses
2. Construct a $(100-\alpha)100\%$ confidence interval for μ
3. Reject $H_0 : \mu = \mu_0$ if μ_0 lies outside the interval
4. State conclusion

The relationship between confidence intervals and one-sided hypothesis tests is a little more complicated. One-sided tests are often accompanied by one-sided confidence intervals $(-\infty, a)$ or (b, ∞). Though we will stick to the more traditional two-sided confidence intervals. You may see either type of confidence interval reported.

To use a two-sided confidence interval to perform a one-sided test: If you have a C% (e.g. 95%) confidence interval and the entire interval lies within the alternative, you may reject the null at a $\frac{1}{2}(100 - C)\%$ (e.g. 2.5%) level.

$H_0: \mu = \mu_0$

$H_A: \mu > \mu_0$



Example 8.1, 8.2 and 8.3 from course text

Example 4

A scientists wishes to detect small amounts of contamination in the environment. To test her measurement procedure, she spiked 12 specimens with a known concentration (2.5 micrograms/l of lead). The readings for the 12 specimens are

1.9	2.4	2.2	2.1	2.4	1.5	2.3	1.7	1.9	1.9	1.5	2.0

(a) Test at level $\alpha = 0.05$,

$$H_0 : \mu = 2.5 \text{ vs. } H_A : \mu \neq 2.5$$

(b) $\alpha = 0.05$

$$H_0 : \mu = 2.5 \text{ vs. } H_A : \mu < 2.5$$

a) two sided test.

μ = true mean concentration.

$H_0: \mu = 2.5$

true mean concentration of scientististsts measurementi is 2.5 μg/l

$$H_A: \underline{\underline{\mu \neq 2.5}} \leftarrow \mu_0$$

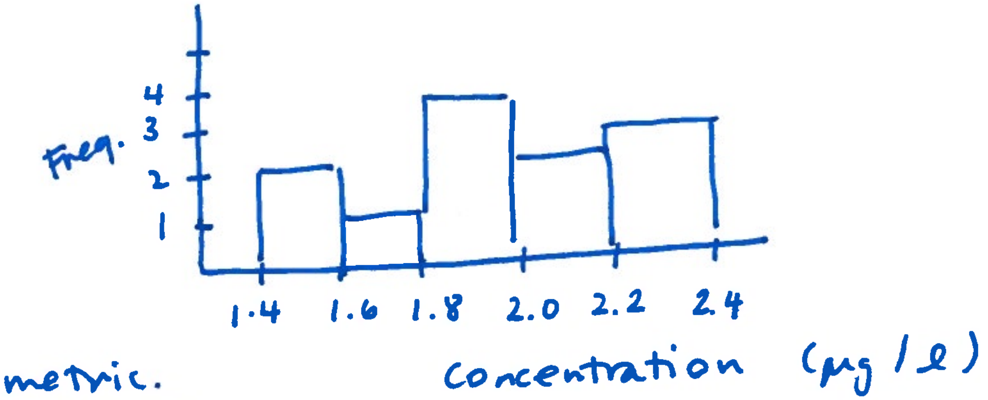
$$\bar{x} = \frac{1.9 + 2.4 + \ldots + 2.0}{12} = \underline{\underline{1.9833}}$$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{(1.9 - 1.9833)^2 + \ldots + (2.0 - 1.9833)^2}{11} = 0.09787879$$

① $n = 12$ small

② σ unknown.

$\rightarrow$ $\underline{\underline{\text{t-test}}}$



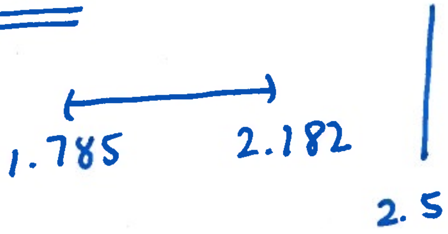
- roughly unimodal, symmetric.
- random sample - representative
- independent.

$\alpha = 0.05$ → 95% CI for μ:

$$\bar{x} \pm t^*_{\underset{\underset{11}{\uparrow}}{n-1},\ \underset{\underset{2.5}{\uparrow}}{\frac{0.05}{2}}} SE(\bar{x}) = 1.9833 \pm \overset{\overset{t_{11,0.025}}{\swarrow}}{2.201} \frac{\sqrt{0.09787879}}{\sqrt{12}}$$

$$= (1.785, 2.182)$$

$H_0: \mu = 2.5$



Since 2.5 μg/ℓ is not in interval

Reject H₀.

Strong evidence that the true mean concentration is not equal to 2.5 μg/ℓ.

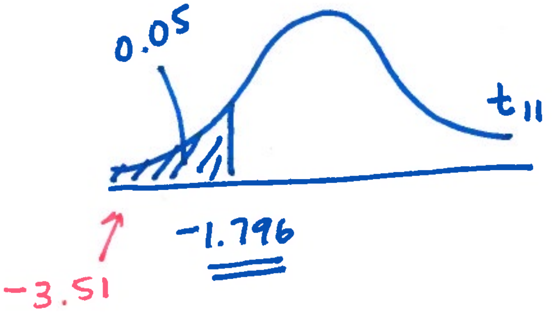
b) Method 1: Critical region method:

$H_0: \mu = 2.5$ vs. $H_A: \mu < 2.5$

$\alpha = 0.05$

Test stat:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.9833 - 2.5}{\sqrt{0.0979}/\sqrt{12}} = -3.51$$



Since −3.51 is under critical region. we reject Ho and conclude there is evidence that the population mean concentration is less than 2.5 μg/ℓ

Method 2: $\alpha = 0.05$

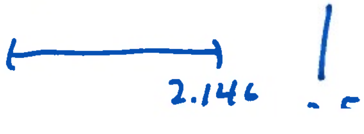
Construct a one sided test using a two-sided interval. $(1 - 2 \times \alpha)100\% = 90\%$ CI.

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 1.9833 \pm 1.796 \frac{\sqrt{0.09787879}}{\sqrt{12}}$$

$$= (1.821, 2.146)$$

$H_0: \mu = \mu_0$ vs.

$H_A: \mu < 2.5$



Reject Ho $\alpha = 0.05$.

Type I and Type II Errors

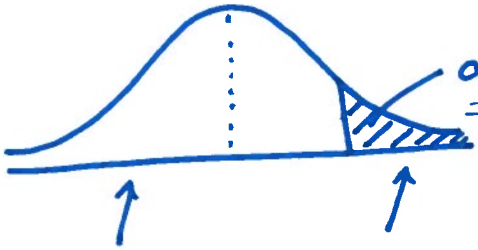
Type I Error is rejecting H_0 when it is actually true
Type II Error is failing to reject H_0 when it is actually false

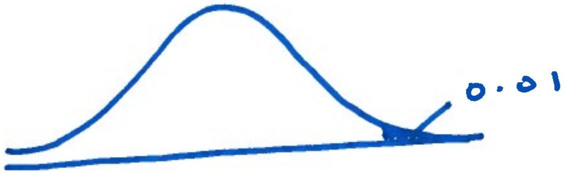
		Truth	
		H_0 is true	H_A is true
Decision	Reject H_0	Type I error	Correct decision
	Do not reject H_0	Correct decision	Type II error

The probability of committing the type I error is α, the significance level of the hypothesis test.
It is the probability of rejecting H_0 when H_0 is true.

$H_0: \mu = \mu_0$ vs. $H_A: \mu > \mu_0$

Null model sample mean:





α too small: ↑ P(type II error)

H_0: patient does not have disease.

H_A: " " " have disease.

→ Type I error: patient healthy, test says patient has disease.

→ Type II error: patient has disease, test says healthy.

increase n, ↓ type I and ↓ type II error.

σ known $\rightarrow$ z ($\underline{\underline{n}}$ large or $\underline{\underline{x}}$ Normal)

$\underline{\underline{\sigma}}$ unknown $\rightarrow$ $\underline{n \text{ small}}$ $\rightarrow$ $\underline{\underline{t}}$ ($\underline{\underline{\text{dist}}}$. $\underline{\underline{\text{Normal}}}$)

$\uparrow$ $\rightarrow$ $\underline{n \text{ large}}$ $\rightarrow$ $\underline{\underline{t}}$ or $\underline{\underline{z}}$

s

$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ ← fixed.

$SE(\bar{X}) = \frac{s}{\sqrt{n}}$ ← random.

Problem 8.6 from course text

Example 5

An automobile manufacturer recommends that any purchaser of one of its new cars bring it in to a dealer for a 3000-mile checkup. The company wishes to know whether the true average mileage for initial servicing differs from 3000. A random sample of 50 recent purchasers resulted in a sample average mileage of 3208 and a sample standard deviation of 273 miles. Does the data strongly suggest that true average mileage for this checkup is something other than the recommended value? ($\alpha = 0.05$)

$\bar{x} = 3208$

$s = 273$.

① pop dist unknown.

② σ unknown.

③ $\underline{\underline{n = 50}}$ large, by CLT

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ approx.

→ random sample of cars.

→ assume cars independent.

Let μ = true average mileage for initial servicing of new cars from manufacturer.

$H_0 : \underline{\underline{\mu = 3000}}$ vs. $H_A : \mu \neq 3000$

$\uparrow$ μ_0

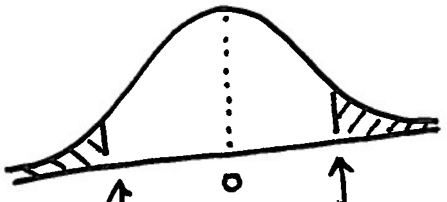
Method 1: <u>critical region.</u>

one-sample t-test

under null, $t_{obs} = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} = \dfrac{3208 - 3000}{273 / \sqrt{50}} = \underline{\underline{5.387}}.$

Critical region:

H_A



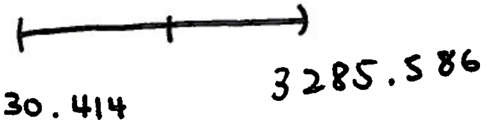
-2.009575 $\quad t_{\underline{49}, 0.025} = \overset{\text{R: qt()}}{\downarrow}\ 2.009575$

t_{obs} falls in critical region. Reject H_0.

Conclude data suggests the true average mileage differs from 3000 miles at 5% sig. level.

<u>Method 2:</u> $H_0 : \mu = 3000$ vs. $H_A : \mu \neq 3000$

$\bar{x} \pm t^*_{n-1} \dfrac{s}{\sqrt{n}} = 3208 \pm 2.009575 \dfrac{273}{\sqrt{50}} = (3130.414,\ 3285.586)$

3000 

Since 3000 does not fall in interval → Reject H_0.

With sufficiently large n,

CLT

$$\bar{x}_1 \overset{\text{approx}}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \bar{X}_2 \overset{\text{approx}}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$$\bar{X}_1 - \bar{X}_2 \overset{\text{approx}}{\sim} N\left(E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2,\ Var(\bar{X}_1 - \bar{X}_2)\right)$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2)$$ independent. $\bar{X}_1, \bar{X}_2$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$SD(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

s_1, s_2

pooled sample variance

$$\underline{\underline{S_p}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Recall:

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

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

Two Sample Problems

Objective: to compare the means of two independent populations e.g. we may want to compare the mean reduction in blood pressure between Drug A and Drug B.

Suppose we draw a random sample from each of the two independent populations: with population means μ_1, μ_2 and population standard deviations σ_1, σ_2.

Our goal is to compare the means μ_1 and μ_2 based on the two independent samples.

Pooled two-sample t-test for difference in means of two independent populations

- The null hypothesis has the form:
  $H_0 : \mu_1 - \mu_2 = \Delta_0$ (i.e. for $\Delta_0 = 0, \;\; H_0 : \mu_1 = \mu_2$)
- The alternative hypothesis takes one of the following three forms (depends on the context):
  - $H_A : \mu_1 - \mu_2 < \Delta_0$ (i.e. for $\Delta_0 = 0, \;\; H_A : \mu_1 < \mu_2$) (left-tailed)
  - $H_A : \mu_1 - \mu_2 > \Delta_0$ (i.e. for $\Delta_0 = 0, \;\; H_A : \mu_1 > \mu_2$) (right-tailed)
  - $H_A : \mu_1 - \mu_2 \neq \Delta_0$ (i.e. for $\Delta_0 = 0, \;\; H_A : \mu_1 \neq \mu_2$) (two-tailed)

Assumptions:

1. Two samples are randomly drawn from their respective population
2. The sampled individuals are independent of each other
3. Both populations are normal or we need reasonably large sample size to validate using the CLT
4. Both population distributions have equal variances $\underline{\underline{(\sigma_1^2 = \sigma_2^2)}}$

Rule of thumb: $\frac{sd_{large}}{sd_{small}} < 2$

Test statistic:

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{SE_{pooled}(\bar{x}_1 - \bar{x}_2)}$$

where

$$SE_{pooled}(\bar{x}_1 - \bar{x}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

with $df = n_1 + n_2 - 2$

and

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Confidence intervals for difference in means of two independent populations

The level C confidence interval for the difference between the two population means, $\mu_1 - \mu_2$, is:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times SE(\bar{x}_1 - \bar{x}_2)$$

Example 8.4 and 8.5 from text

Example 6

A shipyard must order a large shipment of lacquer from a supplier. Besides other design requirements, the lacquer must be durable and dry quickly. A sample of thirty 20-liter cans from Supplier A yields an average drying time of 22.3 minutes and standard deviation of 2.9 minutes. Another supplier, called Supplier B, could also supply the lacquer. A sample of ten 20-liter cans from supplier B yields an average drying time of 20.7 minutes and standard deviation of 2.5 minutes.

(a) Find a 95% confidence interval for $\mu_A - \mu_B$.

(b) Does the data support Supplier B's claim that, on average, its product dries faster than A's? $(\alpha = 0.05)$

a) $$\bar{x}_A - \bar{x}_B \pm t^*_{n_A + n_B - 2} \underbrace{SE(\bar{X}_A - \bar{X}_B)}_{S_{pooled}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

$$S_{pooled}^2 = \frac{(30-1)\, 2.9^2 + (10-1)\, 2.5^2}{30+10-2} = 7.8984$$

$t^*_{38,\, 0.025}$

↑ 95% CI so $\frac{\alpha}{2} = 0.025$ (2 sided)

Use $t_{30,\, 0.025} = 2.042$ (to be conservative.)

$$\underbrace{(22.3 - 20.7)}_{1.6} \pm 2.042 \times \sqrt{7.8984}\sqrt{\frac{1}{30} + \frac{1}{10}}$$

$$= (-0.496,\ 3.696)$$

b) Let μ_A be the mean drying time of all cans produced by A.

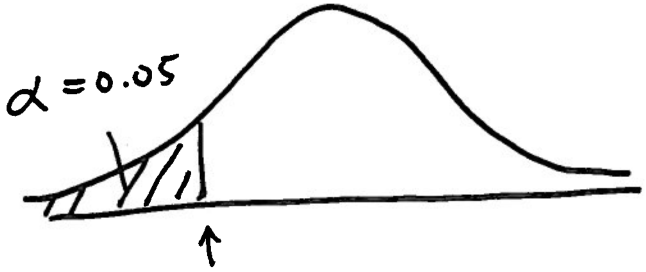
μ_B " " B.

$H_0: \mu_A = \mu_B$ vs. $H_A: \mu_B < \mu_A \rightarrow \underline{\underline{\mu_B - \mu_A}} < 0$ ↑

$\mu_B - \mu_A = 0$ ←

Under null,

$$t_{obs} = \frac{(\bar{x}_B - \bar{x}_A) - 0}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{(20.7 - 22.3) - 0}{\sqrt{7.8984}\sqrt{\frac{1}{30} + \frac{1}{10}}} = -1.56$$



$t_{38,\,0.05} \approx t_{30,\,0.05} = -1.697.$

Since t_{obs} does not fall in critical region we dont reject. Ho.

Not enough evidence to say B's true avg. drying time is sig. smaller than As at 5% level.

Example 8.6 from text

Example 7

Either 20 large machines or 30 small ones can be acquired for approximately the same cost. One large and one small machines have been experimentally run for 20 days with the following results:

$$\bar{y}_{large} = \bar{y}_1 = 31.0, \ s_{large} = s_1 = 2.1$$
$$\bar{y}_{small} = \bar{y}_2 = 22.7, \ s_{small} = s_2 = 1.9$$

Is there statistical evidence in favor of either type of machine? Use $\alpha = 0.05$.

μ_1 = true mean output from large machine

μ_2 = " " " " small "

Test hypothesis: Total output 20 large: $20\mu_1$

30 small $30\mu_2$

$H_0: 20\mu_1 = 30\mu_2 \rightarrow H_0: 20\mu_1 - 30\mu_2 = 0$

$H_A: 20\mu_1 \neq 30\mu_2 \rightarrow H_A: 20\mu_1 - 30\mu_2 \neq 0$

Need to find 95% CI for $20\mu_1 - 30\mu_2$

$$E(20\bar{X}_1 - 30\bar{X}_2) = 20\,E(\bar{X}_1) - 30\,E(\bar{X}_2)$$
$$= 20\mu_1 - 30\mu_2$$

$$20\bar{X}_1 - 30\bar{X}_2 \pm t^*_{df}\ SE$$

$$\nearrow$$

$$\mathrm{Var}(20\bar{X}_1 - 30\bar{X}_2) = 20^2\,\mathrm{Var}(\bar{X}_1) + 30^2\,\mathrm{Var}(\bar{X}_2)$$ assume indep.

$$= 20^2\frac{\sigma_1^2}{n_1} + 30^2\frac{\sigma_2^2}{n_2}$$ $\underline{\underline{\sigma_1^2 = \sigma_2^2}}$ assume

$$= \sigma^2\left(\frac{20^2}{n_1} + \frac{30^2}{n_2}\right)$$

$$= \sigma^2\left(\frac{20^2}{20} + \frac{30^2}{20}\right).$$

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{19\times 2.1^2 + (9\times 1.9^2}{20+20-2} = 4.01$$

$$t^*_{df=38,\ \frac{0.05}{2}} \approx t^*_{\underset{\uparrow}{30},\,0.025} = 2.042$$

conservative.

$$95\%\ \text{CI}: \quad 20\times 31 - 30\times 22.7 \pm 2.042\ S_p\sqrt{\frac{20^2}{20}+\frac{30^2}{20}}$$

$$-61 \pm 2.042\sqrt{4.01}\sqrt{\frac{20^2}{20}+\frac{30^2}{20}}$$

$$= \underline{\underline{(-93.97, -28.03)}}$$

$$H_0: 20\mu_1 - 30\mu_2 = \underset{\nwarrow\ \mu_0}{0}$$

Reject at $\alpha = 0.05$ the hypothesis that both alternatives are equally convenient. We're 95% confident that the true mean output for 20 large machines is between 28.03 and 93.97 <u>lower</u>. than 30 small.

Appears 30 small more convenient.