Ch. 8 - Statistical Modelling and Inference

Population vs. sample, parameter vs. statistic

Population: contains the entire collection of individuals we want to study. The population may be finite or infinite (e.g. Canadian population vs. coin toss)

Sample: subset of individuals selected from the population. Selected randomly

Parameter: characteristic of interest from the population. Value of the parameter is unknown in practice

Statistic: numerical measure of the sample. We use statistics to estimate the unknown population parameter. Due to sampling variability a statistic takes on different values for different samples.

e.g. We use the **sample mean** ($\bar{x}$) to estimate the **population mean** (μ).

We use the **sample standard deviation** (s) to estimate the **population standard deviation** (σ).

We use **Statistical Inference** procedures to make statements about our population of interest using sample data. We will learn:

1. Point Estimation - most likely value for our parameter
2. Confidence Intervals - plausible range for our parameter
3. Hypothesis Testing - Testing if the parameter is equal/not equal to some hypothesized value

ulation with mean μ and variance σ^2.

$$\frac{\sum_{}^{n} X_i}{} \quad 1 \quad \frac{n}{} \quad 1$$

1 One Sample Problems

1.1 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:

"^" estimate

$$E(\hat{\theta}) = \theta$$

- $\bar{x}$ is an unbiased estimator of μ
- s^2 is an unbiased estimator of σ^2

Suppose that $X_1, ..., X_n$ is a simple random sample from a population with mean μ and variance σ^2.

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

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{\sum x_i^2 - 2\sum x_i \bar{x} + n\bar{x}^2}{n-1} \\ &= \frac{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}{n-1} \\ &= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \end{aligned}$$

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

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

$$E(s^2) = E(\frac{\sum x_i^2 - n\bar{x}^2}{n-1}) = \frac{1}{n-1}\left[E(\sum x_i^2) - nE(\bar{x}^2)\right]$$

$$= \frac{1}{n-1} n[E(x_i^2) - E(\bar{x}^2)]$$

$$= \frac{n}{n-1}[\sigma^2 + \mu^2 - (\frac{\sigma^2}{n} + \mu^2)]$$

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

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

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

Notice: $\left[Var(X) = E(X^2) - E(X)^2 \right]$ (σ^2, μ^2)

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

$$Var(\bar{x}) = E(\bar{x}^2) - E(\bar{x})^2$$

$$Var(\sum \frac{x_i}{n}) = E(\bar{x}^2) - E(\bar{x})^2$$

$$\frac{n}{n^2} Var(x_i) = E(\bar{x}^2) - \mu^2 \implies \frac{1}{n}\sigma^2 = E(\bar{x}^2) - \mu^2$$

$$\implies E(\bar{x}^2) = \frac{1}{n}\sigma^2 + \mu^2$$

1.2 Confidence Interval for the population mean μ

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population. Recall that we discussed the sampling distribution of the sample mean.

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 $\bar{x}$ follows approximately $N(\mu, \frac{\sigma^2}{n})$. CLT
- If the underlying distribution of x is Normal, the sampling distribution of $\bar{x}$ is Normal and n need not be large.

In most situations, the population standard deviation σ is unknown.
So we estimate the population standard deviation σ using the sample standard deviation s. We use the **standard error**

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

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

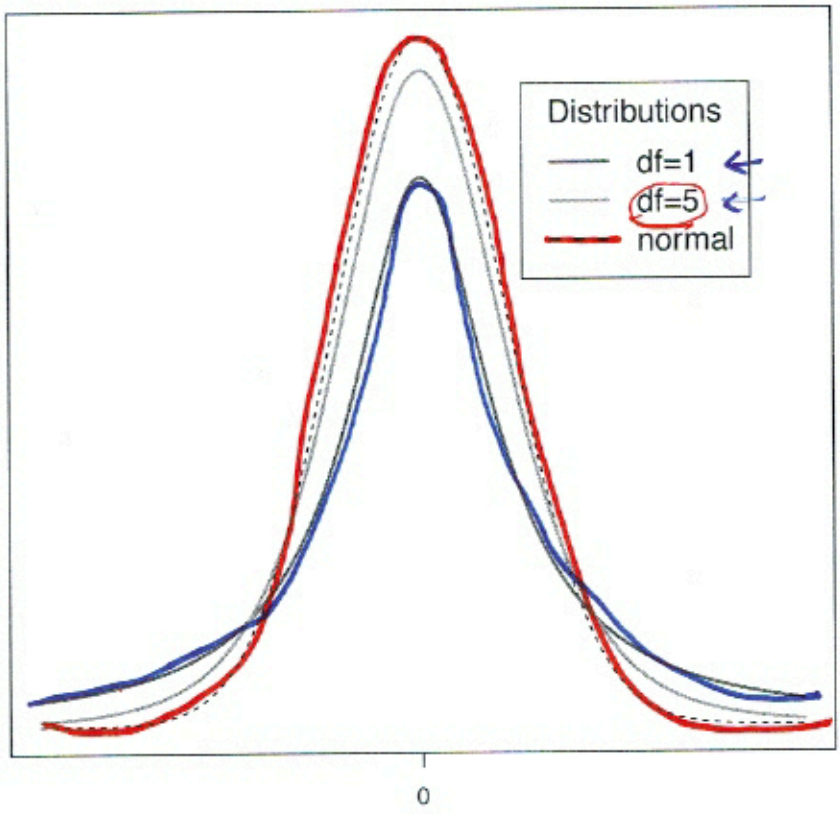
For large samples, s is a reliable estimate for σ. For small samples, s is likely a poor estimate for σ. The standard error has an extra source of variation from the sampling variability of s. Consider an independent, random sample from a certain population with mean μ and standard deviation σ. For a 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 $\frac{\bar{x}-\mu}{SE(\bar{x})}$) 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.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The sampling distribution of t is called the **Student's t-model**. This distribution has thicker tails than the standard Normal model and the shape of the distribution changes with sample size.

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.

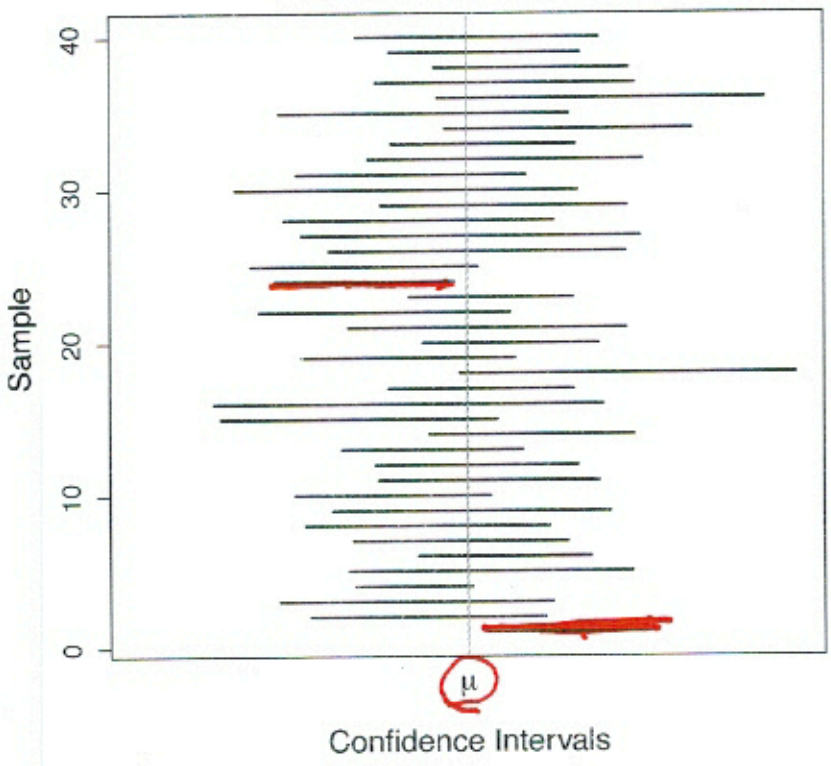


A **confidence interval** is used to describe the amount of uncertainty associated with a sample estimate of a population parameter. It consists of a range of values that act as plausible estimates of the population parameter. In general, a **confidence interval** takes the form:

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

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

A 95% confidence level means that in a very large number of samples, 95% of the intervals would contain the parameter.



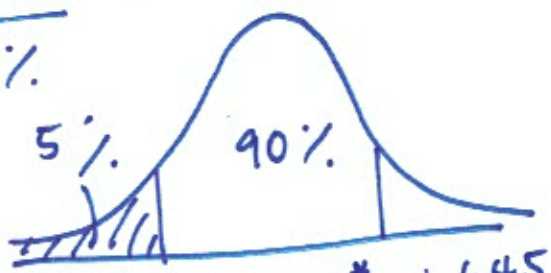
One-sample z confidence interval

- 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:

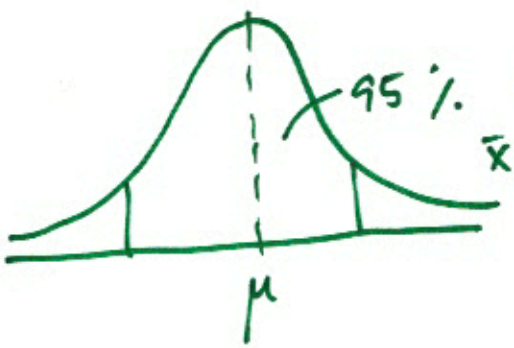
$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

estimate ± margin of error.

z^*	1.645	1.960	2.576
C	90%	95%	99%



↑ z value corresponds to 95%



One-sample t confidence 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 C for μ is computed using:

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

$$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

ME

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 (e.g. $n \geq 30$), t_{n-1} and $N(0, 1)$ are almost identical and we can use z as the test statistic.
For small samples (e.g. $n < 30$), t_{n-1} and $N(0, 1)$ are different and we use t as the test statistic.

Conditions for constructing confidence intervals:

- sample is randomly drawn from the population
- sampled values are independent
- the random sample must come from a (nearly) normal population distribution or sample size is large.

Summary of when to use z or t

Population Distribution	Population Variance	Sample size	t or z
Normal	known	any size	z
Normal	unknown (estimate σ by s)	large	z
		small	t
Unknown (for large samples we can assume a normal dist by the CLT)	known	large	z
	unknown (estimate σ by s)	large	z

Example 1. Assume that the helium porosity (in percentage) of coal samples is normally distributed with true standard deviation 0.75.

(a) Compute the 95% confidence interval for the true mean porosity if the average porosity for a random sample of 20 specimens was 4.85.

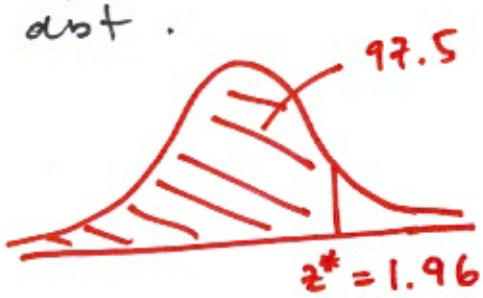
(b) What sample size is necessary for the margin of error to be 0.2 with 99% confidence?

Solution:

a) - population is normally dist.

- dist σ is known.

$\rightarrow$ z - interval.



$\bar{x} = 4.85$

$n = 20$

$\sigma = 0.75$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

$$4.85 \pm 1.96 \times \frac{0.75}{\sqrt{20}}$$

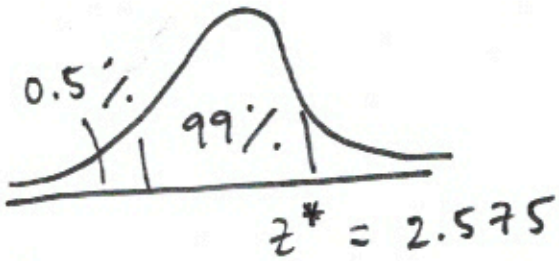
$$= 4.85 \pm 0.33$$

$$= (4.52, 5.18)$$

we are 95% confident the true mean porosity is between 4.52 and 5.18.

b) $mE = 0.2$

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



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

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

$$n = \left(\frac{2.575 \times 0.75}{0.2} \right)^2 = 93.24 \rightarrow n = 94$$

(round up)

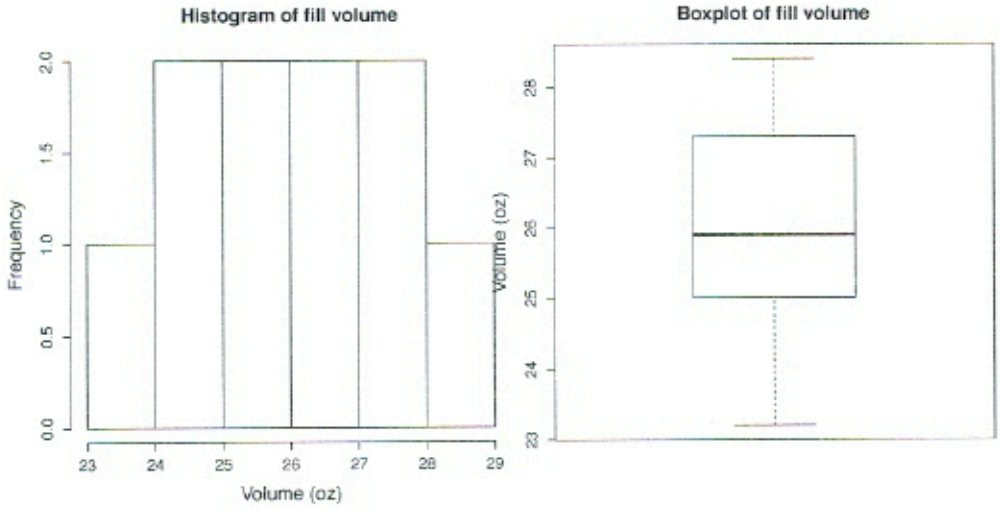
Example 2. A machine is used to fill containers with a liquid product. A random sample of 10 containers is selected and the contents are shown below.

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.



Solution:

a) histogram, + boxplot → unimodal + symmetric.

small n
unknown σ. → use t

$$\bar{x} = \frac{\sum x_i}{n} = \frac{260}{10} = 26$$

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

$$= \frac{1}{9}[6783.76 - 10 \times 26^2$$

$$= 2.64$$

b) $\bar{x} = 26$ oz

$s^2 = 2.64$ oz

$$SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{1.62}{\sqrt{10}} = 0.514$$

$$ME = t^* \frac{s}{\sqrt{n}} = 1.833 \times 0.514 = 0.942$$

t^*: $t_{9, \frac{\alpha}{2}} = \frac{0.1}{2} = 0.05$

1.833: t-table.

c) 90% CI for μ.

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 26 \pm 0.942$$

$$= (25.06, 26.942)$$

90% confident that the true mean fill volume is between 25.06 oz and 26.942 oz.

$df = n - 1 = 9$

Finer Details of Confidence Intervals

We can say: "We are 90% confident that the true mean fill volume is between 25.06 and 26.94 ounces."

We can't say the following: (What might be wrong with each of the following statements?)

- the true mean fill volume is between 25.06 and 26.94 ounces

10% of intervals we make will not contain μ. We need to attribute confidence level.

- the true mean fill volume is 26oz, 90% of the time

→ implies the true mean fill volume varies. CIs vary around mean.

- 90% of all containers have between 25.06 and 26.94 oz

Should be about MEAN not individual containers.

- we are 90% confident the sample mean fill volume is between 25.06 and 26.94 oz

1.3 Testing of Hypotheses about μ

Hypothesis Testing Basics and Definitions

- In statistics, a **hypothesis** is a statement or claim about a parameter. We test this claim with a hypothesis test
- The **null hypothesis** (H_0) is a statement about the value of the population parameter

  H_0 : population parameter $=$ 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.

  Three possible forms:

  H_A : population parameter $\neq$ some hypothesized value (two-tailed)
  H_A : population parameter $>$ some hypothesized value (one-tailed)
  H_A : population parameter $<$ some hypothesized value (one-tailed)

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

$$\begin{array}{lll} H_0 : \mu = \mu_0 & \text{vs.} & H_A : \mu \neq \mu_0 \\ H_0 : \mu = \mu_0 & \text{vs.} & H_A : \mu > \mu_0 \\ H_0 : \mu = \mu_0 & \text{vs.} & H_A : \mu < \mu_0 \end{array}$$

A certain machine is meant to produce metal rods of length 3m. After several years of operation, it is suspected that the rod lengths have decreased and the machine needs to be serviced. To determine whether there is sufficient evidence to support the suspicion, we could take a random sample of, say, 100 rods. We could conduct a hypothesis test about our true population mean using the average length from our sample. Suppose our sample has an average rod length of 2.75m. Does this difference indicate that our machine needs servicing?

In this example about metal rods, the hypothesis test is carried out on the true mean length of rods produced by the machine. Let μ = the true mean length (in m) of rods produced by the machine.

$$\begin{array}{l} H_0 : \mu = 3 \\ H_A : \mu < 3 \end{array}$$

μ_0

we are interested in testing if population mean has decreased.

We begin by assuming the null model is correct and use the sample data and our knowledge of statistical theory to decide whether the sample data supports the null hypothesis or provides sufficient evidence against it.

- A **test statistic** is a statistic calculated using the data and later used to decide whether to reject the null hypothesis. We will assess how unusual our observed difference is if the null hypothesis is true by computing a test statistic.
- 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. It is the probability of rejecting H_0 when it H_0 is true (think of it as P(rejecting H_0 | H_0 is true) = α).
- The **critical region** consists of outcomes that are very unlikely to occur if the null hypothesis is true.
- A large value for the test statistic shows that the obtained mean difference is more than would be expected if there is no treatment effect. If it is large enough to be in the critical region, we reject the null hypothesis.
- If the mean difference is relatively small, then the test statistic will have a low value. In this case, we conclude that the evidence from the sample is not sufficient, and the decision is fail to reject the null hypothesis.

How to perform 2-sided tests $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$

Method 1: Using Confidence Intervals

(a) Set up null and alternative hypotheses

(b) Construct a confidence interval for μ

(c) Reject $H_0 : \mu = \mu_0$ if μ_0 lies outside the interval

(d) State conclusion

Method 2: Critical Region Method

(a) Set up null and alternative hypotheses

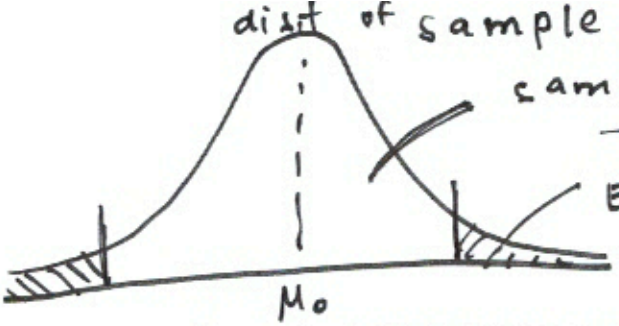
(b) Decide on the level of significance (α)

(c) Compute the value of the test statistic

(d) Find the approprate critical values

(e) Reject H_0 if the test statistic falls in the rejection region

(f) State conclusion



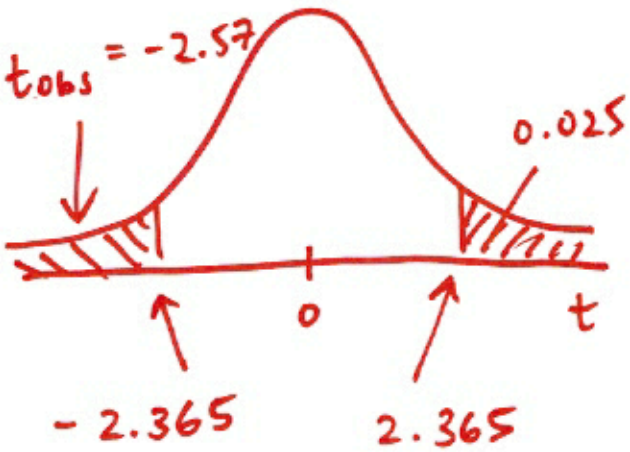
Example 3. Shown using t-distribution (z is also the same).
Say we want to test $H_0 : \mu = 8.3$ vs. $H_A : \mu \neq 8.3$ at $\alpha = 0.05$ significance level and $\bar{x} = 5$, $s = 3.63$, $n = 8$

Test statistic:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5 - 8.3}{3.63/\sqrt{8}} = -2.57$$

We look up $t_{n-1,\alpha/2} = t_{7,0.05/2} = 2.36$. If t_{obs} falls in the critical region, we reject H_0.
Picture:

we look up $t_{n-1, \frac{\alpha}{2}}$



$\alpha = 0.05$

$t_{7, 0.025} = 2.365$

from table

We check if t_{obs} falls in critical region.

Here we reject H_0 and conclude the observed difference is significant or real.

How to perform 1-sided tests

It is much easier to use the critical region method for 1-sided tests.

Example 4. Say we want to test $H_0 : \mu = 8.3$ vs. $H_A : \mu < 8.3$ at $\alpha = 0.05$ significance level and $\bar{x} = 5$, $s = 3.63$, $n = 8$

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = -2.57$$

We look up $t_{n-1,0.05} = 1.89$ (And we use $-t_{n-1,0.05}$ since the critical region is on the left side). $t_obs = -2.57$ falls in the critical region, so we reject H_0.

Picture:

test statistic same as before $t_{obs} = -2.57$.

$t_{n-1,0.05} = t_{7,0.05}$

We look up $t_{7,0.05}$

$t_{obs} = -2.57$ falls in critical region so we reject the null hypothesis.

Example 8.1, 8.2 and 8.3 from course text

Example 5. 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

$\bar{x} = 1.9833$

$s^2 = 0.09782879$

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

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

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

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

Solution:

μ not $\bar{x}$

a) Two sided test

$H_0 : \mu = 2.5$ vs $H_A : \mu \neq 2.5$

H_0: the true mean contamination is 2.5 µg/l

H_A: true mean contamination is different from 2.5 µg/l

Let μ be the true mean of the scientist's measurements.

95% confidence interval for μ.

$$\bar{x} \pm t^*_{11,0.025} \cdot SE(\bar{x})$$

$$1.9833 \pm 2.201 \times \frac{\sqrt{0.09787879}}{\sqrt{12}}$$

μ_0 = (1.785, 2.182)

Since 2.5 µg/l is not in the interval (1.785, 2.182) we reject H_0.

There is statistical evidence that our population is not equal to 2.5 µg/l

b) $H_0: \mu = 2.5$

vs. $H_A: \mu < 2.5$ one sided test.

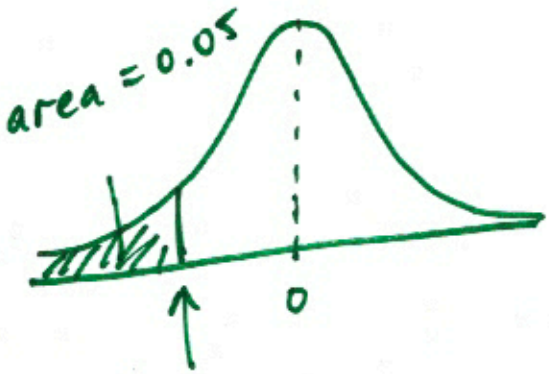
$\alpha = 0.05$

Under H_0,

test statistic

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.9833 - 2.5}{\sqrt{0.09787879}/\sqrt{12}}$$

$$= -3.51$$



$= -1.769$

We look up

$t_{11, 0.05} = 1.796$

and since $H_A: \mu < 2.5$

our critical region is the area to the left of $t = -1.769$

Since $t_{obs} = -3.51$ is in the critical region. Reject H_0 and conclude there is statistical evidence that the population mean concentration is less than 2.5 µg/l

Problem 8.2 from course text

Example 6. 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.

Thought process:

① σ^2 is unknown

② sample size is small (must use t)

③ Population normally distributed

a) $\bar{x} = \frac{\sum x_i}{n}$ x_i repair time for ith instrument.

$\bar{x} = \frac{212 + \ldots + 250}{10} = 227.60$ hours.

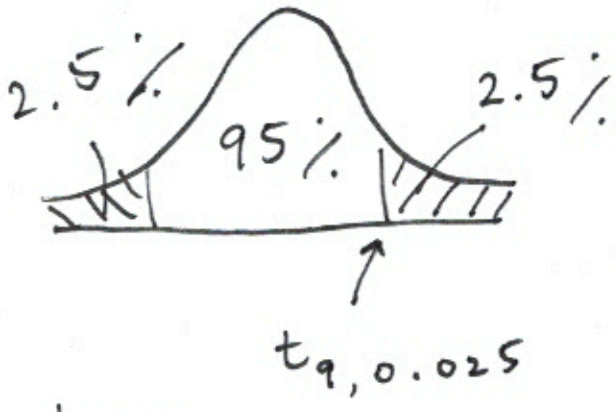
$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \left(\frac{(\sum x_i)^2}{n}\right)}{n-1}$

$$= \frac{546608 - \frac{2276^2}{10}}{9} = 3176.71$$

b) 95% CI for μ. $df = n - 1 = 9$

$\bar{x} \pm (t) SE(\bar{x})$

$227.6 \pm 2.262 \sqrt{\frac{s^2}{n}}$



(187.28, 267.92) hours.

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

c) $H_0: \mu = 200$ ← μ_0

$H_A: \mu > 200$

Note: the worker claims his avg. repair time is ≤ 200 hours.

One sided test.

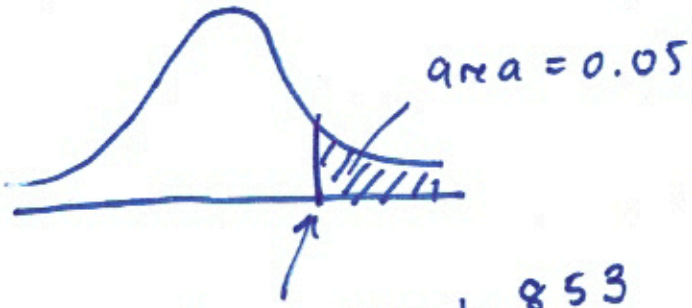
Under H_0,

test stat.

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{227.60 - 200}{17.823}$$

$$= 1.528$$

Let's assume $\alpha = 0.05$



$t_{9, 0.05} = 1.853$

$t_{obs} = 1.528$ does not lie within the critical region. Therefore we do not reject the null hypothesis.

Problem 8.6 from course text

Example 7. 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?

Solution:

Thought process:

① sample size is large ($n = 50$)

② pop. dist is unknown.

③ we have a sample variance (273^2)

We use one sample z test

Let μ be the true average mileage.

$H_0: \mu = 3000$ ← μ_0

$H_A: \mu \neq 3000$

By CLT,

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

estimate this with s^2/n

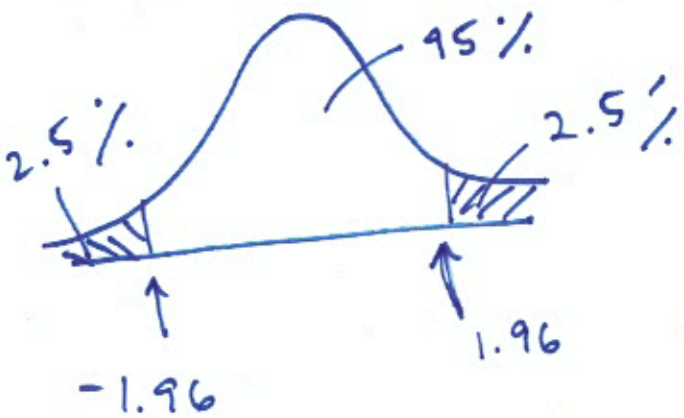
Under H_0,

test stat

$$z_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3208 - 3000}{273/\sqrt{50}}$$

$$= 5.387$$

Critical region (say $\alpha = 0.05$)



Since z_{obs} falls within critical region we reject H_0.

Conclude the true average mileage differs from 3000 miles based on our data.

Example 8. Nielsen, a global information and measurement company, provides insights and data about what people watch, listen to and buy. In 2011, Nielsen reported that U.S. Internet users spent an average of 8.3 hours per month on Facebook.

(a) Suppose we want to determine a 95% confidence interval for the Canadian average and we draw the following random sample from the Canadian population of internet users.

5, 6, 0, 4, 11, 9, 2, 3 ← $\bar{x} = 5$, $s = 3.63$

Find the 95% CI for the average time spent on Facebook among Canadian internet users.

(b) Using your answer in part (a), test at a 5% signficance level if the Canadian average is different from the U.S. average.

(c) Repeat part b, but this time, we wish to test whether the Canadian average is smaller than the U.S. average (at $\alpha = 0.05$).

a) Thought process:

Small sample size of 8.

Can calculate sample mean + sd

→ we dont know pop. variance..

In order to use t, we need to assume population distribution is normal.

$$\bar{x} = \frac{40}{8} = 5 \qquad s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{8}}{n-1}$$

$$= 13.15$$

$$s = 3.625$$

95% CI for μ.

$$\bar{x} \pm t_{7,0.025} \frac{s}{\sqrt{n}}$$

3.63

$t_{7,0.025}$: 2.365

$n = 8$

$$= (2.0, 8.0)$$

b) Let μ be the Canadian population average time on Facebook.

$H_0 : \mu = 8.3$ ← μ_0

$H_A : \mu \neq 8.3$

From part a) we calculated a 95% CI for μ. (2.0, 8.0)

We can use this CI to draw a conclusion about the above hypothesis at $\alpha = 0.05$

CI: (2, 8)

μ_0

Since 8.3 is not in the interval.

We reject H_0 at 5% level.

Conclude Canadian average is significantly different from American average

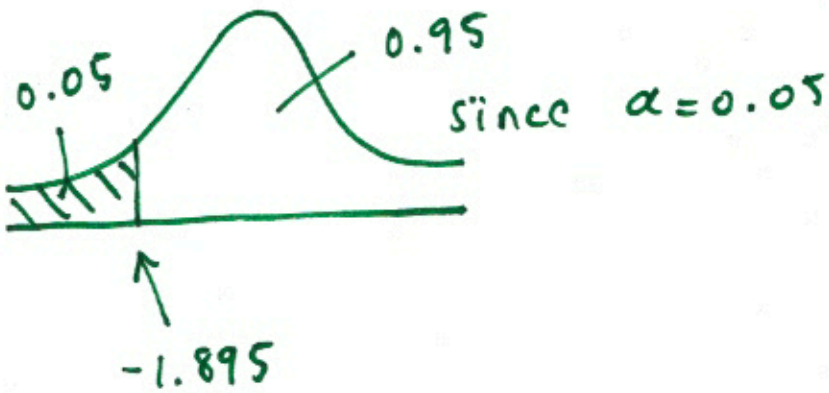
c) $H_0: \mu = 8.3$ ← μ_0

$H_A: \mu < 8.3$ ← μ_0

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5 - 8.3}{3.63/\sqrt{8}} = -2.57$$

we compare $t_{obs} = -2.57$ with

$df = 7$

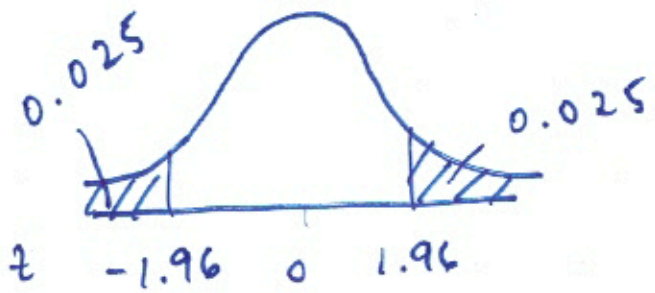


Since $t_{obs} = -2.57$ falls under critical region we reject H_0 and conclude that the Canadian population average is smaller than the U.S.

A side

Two tailed z-test

$\alpha = 0.05$



$H_A: \mu \neq \mu_0$

One tailed z test

$H_A: \mu > \mu_0$

$z_{obs} = 1.8$ → H_0 would be rejected in one sided test but not in the 2 sided test.

1.4 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.

H_0: innocent person

Type I error:
Innocent person falsely convicted.

Type II error
criminal freed

ex. testing disease

Null: No disease
H_A: have disease

Type I: healthy patient but test says they have disease.

Type II: Patient unhealthy, test says no disease

2 Two Sample Problems

2.1 Pooled 2-sample t-tests and confidence intervals

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:

- $x_1^1, x_2^1, ..., x_{n_1}^1$ from a population with mean μ_1 and standard deviation σ_1
- $x_1^2, x_2^2, ..., x_{n_2}^2$ from a population with mean μ_2 and standard deviation σ_2

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 ($\sigma_1^2 = \sigma_2^2$)

Sampling distribution for difference in means of two independent populations

- To estimate $\mu_1 - \mu_2$, we use $\bar{x}_1 - \bar{x}_2$
- We need to calculate an unbiased estimator for the common variance based on s_1^2 and s_2^2

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

$$SE_{pooled}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2}} = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

with $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

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

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

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2)$$

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

because $\bar{X}_1$ and $\bar{X}_2$ indep.

$$SD(\bar{X}_1 - \bar{X}_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

sufficiently large n_i,

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ approx by CLT}$$

$$\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Hence $\bar{X}_1 - \bar{X}_2 \overset{\text{approx}}{\sim} N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

Example 8.4 and 8.5 from text

Example 9. 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) Test if there is sufficient evidence to dispute Supplier B's claim that, on average, its product dries faster than A's. ($\alpha = 0.05$)

a) $(\bar{x}_A - \bar{x}_B) \pm t_{n_A+n_B-2}\, SE$

$SE = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$

$s_p^2 = \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}$

$= \frac{29 \times 2.9^2 + 9 \times 2.5^2}{30+10-2} = 7.8984$

$t_{30+10-2,\,0.025} = t_{38,\,0.025}$

↑ 95% CI so use 0.025

df = 38 not entry in table

so use $t_{30,\ 0.025} = 2.042$

↑ conservative, use smaller df

$$(22.3 - 20.7) \pm 2.042\ s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$22.3 - 20.7 = 1.6$

$s_p = \sqrt{7.898}$

$n_A = 30$

$n_B = 10$

$(-0.496,\ 3.696)$

b) $H_0: \mu_A = \mu_B$ rewrite $H_0: \mu_A - \mu_B = 0$

$H_A: \mu_A < \mu_B$ $H_A: \mu_A - \mu_B < 0$

μ_A be the true mean drying time of supplier A.

μ_B " " B.

Under the null hypothesis,

test stat

$$t_{obs} = \frac{(\bar{x}_A - \bar{x}_B) - 0}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

$s_p = \sqrt{7.8984}$ (from part a)

$$= \frac{(22.3 - 20.7) - 0}{\sqrt{7.8984}\sqrt{\frac{1}{30} + \frac{1}{10}}} = 1.56$$

$-t_{38, 0.05} = -1.697$

$df = n_A + n_B - 2$

use $df = 30$ since 38 not listed

We do not reject H_0.

There is no statistical difference between means.

Example 8.6 from text

Example 10. 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:

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

← error in text.

22.7 error in text

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

$20\mu_1$ = 20 large

$30\mu_2$ = 30 small

$\bar{x}_1 = 31$ $\quad$ $\bar{x}_2 = 22.7$

$s_1 = 2.1$ $\quad$ $s_2 = 1.9$

$n_1 = 20$ $\quad$ $n_2 = 20$

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

$$\text{Var}(20\bar{x}_1 - 30\bar{x}_2) = 20^2 \text{Var}(\bar{x}_1) + 30^2 \text{Var}(\bar{x}_2)$$

assume $\bar{x}_1, \bar{x}_2$ indep.

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

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

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

$$\cdot SE = s_p \sqrt{\frac{20^2}{20} + \frac{30^2}{20}}$$

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

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

$$= \frac{19 \times 2.1^2 + 19 \times 1.9^2}{38} = 4.01$$

$$t_{38,0.025} \simeq t_{30,0.025} = 2.042$$

Hence 95% CI for $20\mu_1 - 30\mu_2$

$$(20\bar{x}_1 - 30\bar{x}_2) = 20 \times 31 - 30 \times 22.7 = -61$$

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

SE

$t_{30,0.025}$ s_p

$$= (-93.97, -28.03)$$

hypothesized value

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

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

Therefore we reject H_0 at $\alpha = 0.05$ because 0 is not in the interval.

Appears 30 small machines is more convenient.