Stat 23400 Lecture 14: Hypothesis Testing for the Mean and Proportion

Outline for Today

- Topics:
  - Hypothesis Testing for the Mean (MMSA 9.2 and OIS 7.1)
  - Hypothesis Testing for Proportion (MMSA 9.3 and OIS 5.3)
- Lab08: Inference for numerical data

Hypothesis Testing for the Mean: Z Test

Old Faithful Dataset

- Old Faithful is a geyser in Yellowstone National Park that is known for erupting approximately once every hour.
- That is, it is claimed that the average waiting time for Old Faithful is 60 minutes.
- We want to see if data support this claim.

Data consist of two variables

- `duration` Eruption time in mins.
- `waiting` Waiting time to this eruption (in mins).

Old Faithful Dataset

```
library(tidyverse) ## for glimpse() function
library(MASS) ## contains geyser dataset
data("geyser")
glimpse(geyser)
```

```
Observations: 299
Variables: 2
$ waiting  <dbl> 80, 71, 57, 80, 75, 77, 60, 86, 77, 56...
$ duration <dbl> 4.017, 2.150, 4.000, 4.000, 4.000, 2.0...
```

```
waiting <- geyser$waiting
```

Hypotheses

Formulating the hypotheses is the always first step in any testing scenario. Using these data, we wish to decide between one of two hypotheses:

- H_0: $\mu = 60$, i.e., the mean waiting time μ for Old Faithful is 60 minutes.
- H_A: $\mu \neq 60$, i.e., the mean waiting time μ for Old Faithful is not 60 minutes.

Wrong Ways to State H_0 and H_A

H_0 and H_A are **ALWAYS** stated in terms of population parameters, not sample statistics.
Neither

$$H_0 : \bar{x} = 60, \quad H_A : \bar{x} > 60$$

nor

H_0 : Waiting time in the sample is 60 minutes on average

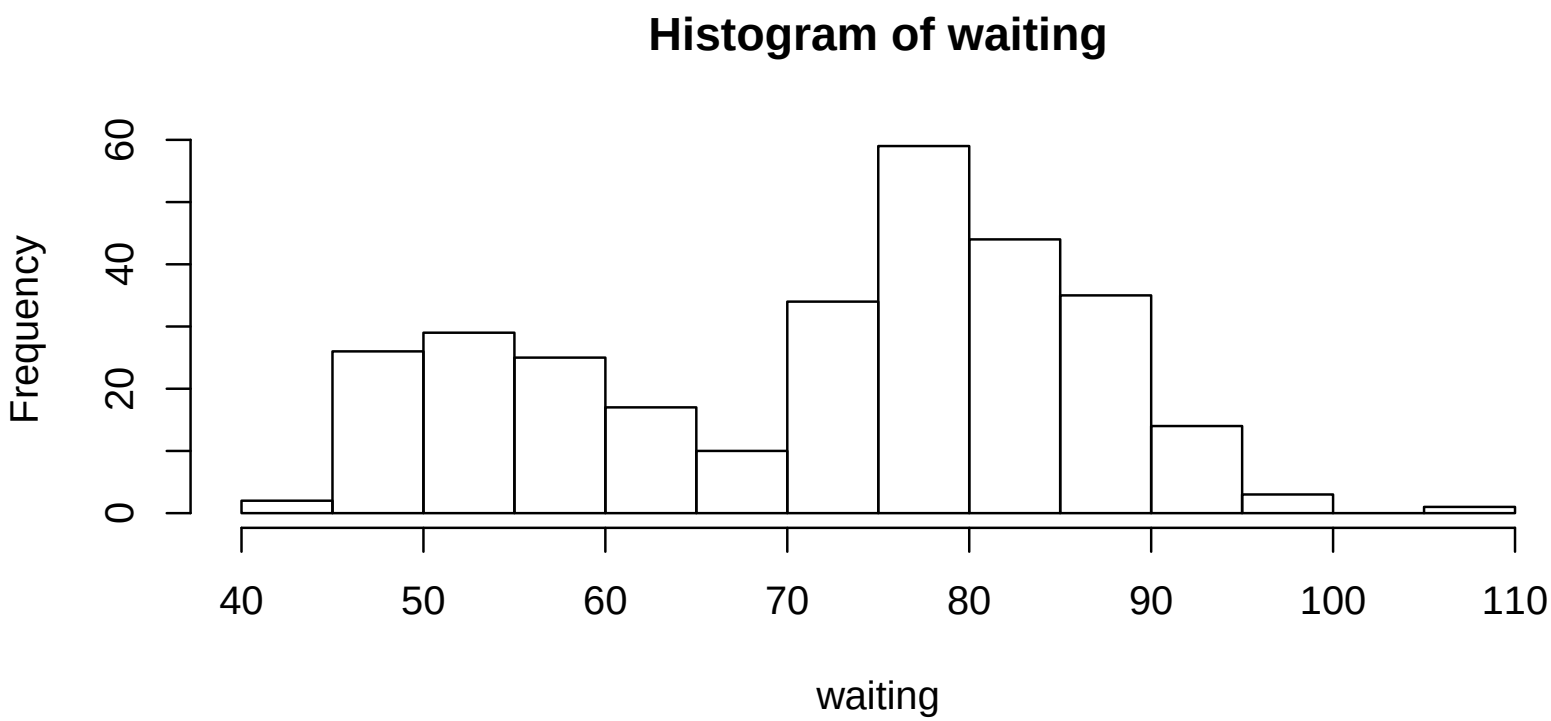
H_A : Waiting time in the sample is 72.3 on average.

is correct.
The correct statements should be

$$H_0 : \mu = 60, \quad H_A : \mu \neq 60.$$

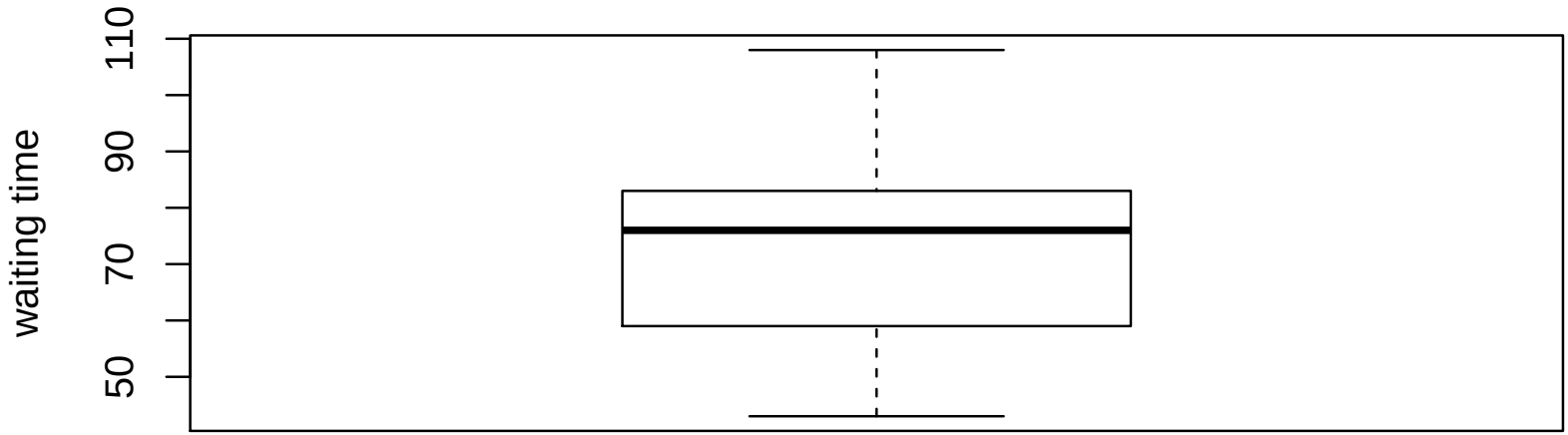
Also please **clearly specify what μ. is.**
e.g., μ is the expected waiting time for the old faithful to erupt.

```
hist(waiting)
```



Exploratory Data Analysis II

```
boxplot(waiting, ylab = "waiting time")
```



Exploratory Data Analysis III

```
summary(waiting)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    59.0    76.0    72.3    83.0   108.0

sd(waiting)

13.89
```

Conclusion from Exploratory Data Analysis

- Exploratory data analysis suggests $\mu \neq 60$, but again, this might be due to random variation.
- We need some formal way to evaluate the unlikeliness of the data we observe under H_0.

Old Faithful — Test Statistic

In the Old Faithful data, $n = 299$, $\bar{x} = 72.3$, $s = 13.89$.

- What is the sampling distribution for $\bar{X}$ under H_0?
- How to tell how unusual the observed sample mean $\bar{x} = 72.3$ is relative to its hypothesized value $\mu_0 = 60$? (What statistic to use?)
- What is the observed value for the Z-statistic?

Calculating *p*-value

- How to calculate the **p-value**?

Calculating *p*-value

```
xbar <- mean(waiting)
s    <- sd(waiting)
n    <- length(waiting)
z    <- (xbar - 60) / (s / sqrt(n))
2 * pnorm(-abs(z))
```

```
[1] 4.837e-53
```

Conclusion: Old Faithful Data

- Since p-value is much low (much lower than 5%) we strongly reject H_0.
- The data provides strong evidence that Old Faithful does not on average erupt once an hour.
- The difference between the null value of an hour and observed sample mean of 72.3 minutes is not due to chance or sampling variability.

General Form of Z Test for the Mean

One-Sided Hypothesis Test for the Mean

- If we wanted to know whether the data provide convincing evidence that the average waiting time is more than an hour, the alternative hypothesis would be different.

$$H_0 : \mu = 60$$
$$H_A : \mu > 60$$

- In this case, a sample mean $\bar{X}$ far below 60 would not be evidence in favor of H_A (only those far above 60). Hence the p-value would be the **one-tail** probability.

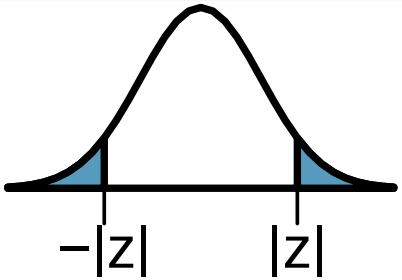
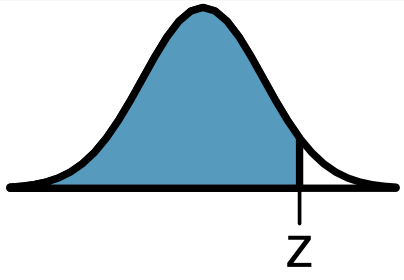
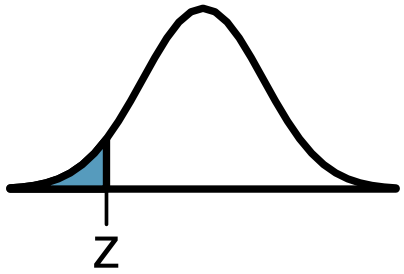
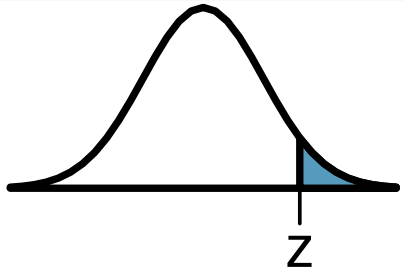
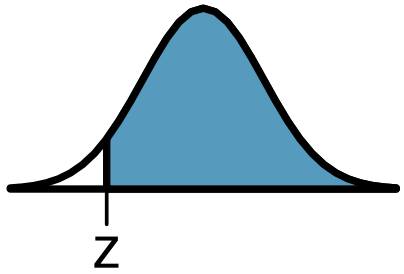
$$\begin{aligned} p\text{-value} &= P(Z \geq z_{obs}) \\ &= P(Z \geq 15.31) \end{aligned}$$

One-Sided v.s. Two Sided Tests

The z-statistic for testing $H_0 : \mu = \mu_0$ is $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$.

Under H_0, $Z \sim N(0, 1)$

The p-value depends on H_A.

	Two-sided test	One-sided test	
H_A	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
P-value		 	 

The bell curve above is the standard normal curve. We reject H_0 when P-value $< \alpha$.

Conclusion when the P-value is Low

When the P-value is lower than the significance level, we say

- The H_0 is rejected, with p-value $=$.
- There is strong evidence that the waiting time for Old Faithful to erupt is not an hour on average (H_A is true).
- The waiting time for Old Faithful to erupt is significantly not an hour.

We don't say

- The H_A is accepted.
- We fail to reject H_A.

Conclusion when the P-value is Not Low

When the P-value exceeds the significance level, we say

- We fail to reject H_0.
- No strong evidence that the waiting time is not an hour for Old Faithful to erupt on average (H_A is true)
- The mean waiting time for Old Faithful to erupt is not significantly different from an hour.

We don't say

- the H_0 is accepted.
- we fail to accept H_A.
- there is strong evidence that H_0 is true — because we might have made a Type 2 error, and the chance of making a Type 2 error is not controlled, which can be quite big.

Z Test for the Mean: Conditions

As CLT is used in the hypothesis test above, we need to check the same conditions as we construct confidence intervals for the population mean.

- Observations must be **independent**.
  - Use your knowledge to judge if the data might be dependent.
- The population distribution should not be extremely skewed.
- In the z-statistic $= \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, if the unknown population SD σ is replaced with the sample SD s, we need to further check that
  - sample size cannot be too small (at least 30).
  - no outliers & not too skewed $\Rightarrow$ Check the histogram of data!

Recap: Hypothesis Testing for a Population Mean

1. Set the hypotheses
   - $H_0 : \mu = \mu_0$
   - $H_A : \mu <$ or $>$ or $\neq \mu_0$
2. Check assumptions and conditions
   - Independence
   - Normality: nearly normal population or $n \geq 30$, no extreme skew – or use the t distribution (Section 5.1)
3. Calculate a **test statistic** and a p-**value** (draw a picture!)

$$Z = \frac{\bar{X} - \mu_0}{SE}, \text{ where } SE = \frac{\sigma}{\sqrt{n}}$$

4. Make a decision
   - If p-value $< \alpha$, reject H_0
   - If p-value $> \alpha$, do not reject H_0

Hypothesis Testing for the Mean: T Test

What if σ is Unknown and Sample size is not big?

- For small n, recall that if $X_1, X_2, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma)$, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a **t-distribution** with $\mathbf{n-1}$ **degrees of freedom.**

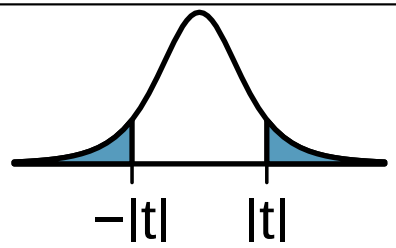
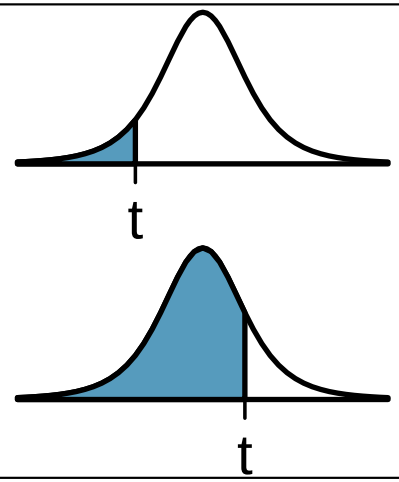
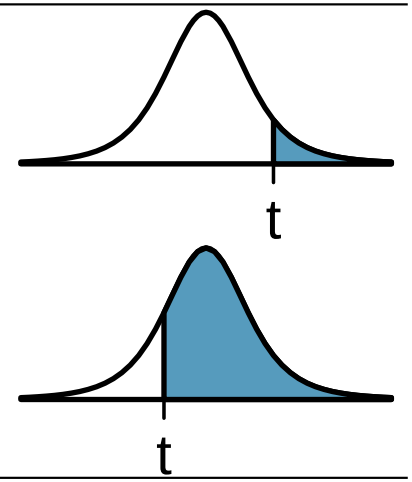
- Similarly as t-based C.I. for the mean, we will derive t-test for the mean.

One-sample t Test of a Population Mean

Similar to the z test, the t-statistic for testing $H_0 : \mu = \mu_0$ is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} .$$

Under H_0, $T \sim t_{n-1}$.
The p-value depends on H_a.

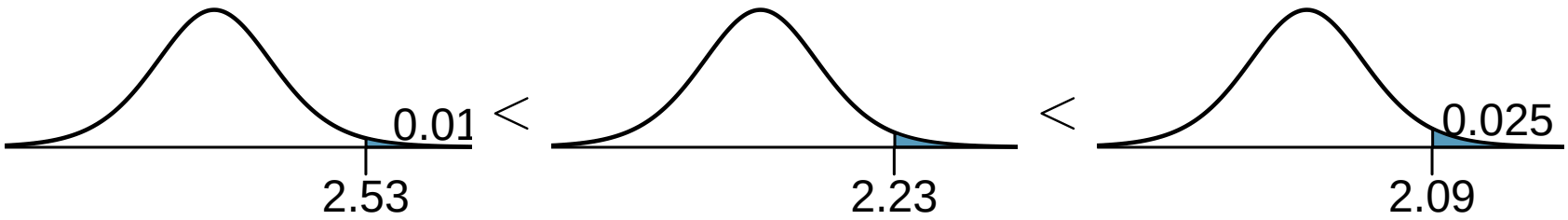
H_a	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
p-value			

The bell curve above is the t-curve with df $= n - 1$, not the normal curve. Then we reject H_0 when P-value $< \alpha$.

How to Use the t-Table to Find p-Values?

Example 1. Testing $H_0 : \mu = 10$ vs. $H_A : \mu > 10$, sample size $n = 21$, the t-statistic is 2.23.

one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81

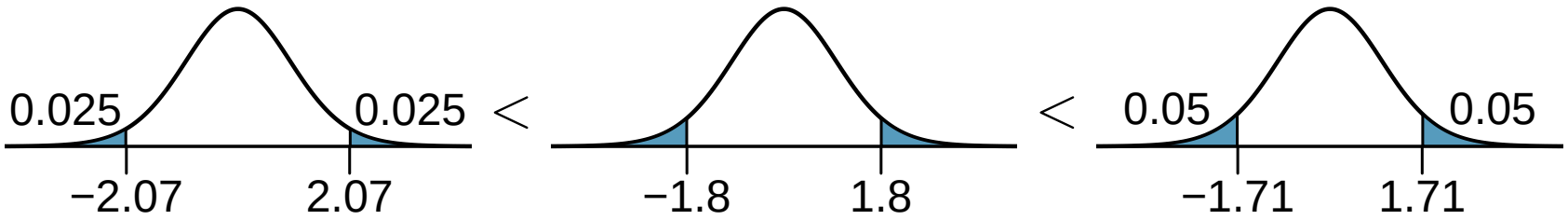


One-sided p-value is between 0.025 and 0.01.

How to Use the t-Table to Find p-Values?

Example 2. Testing $H_0 : \mu = 60$ vs. $H_A : \mu \neq 60$, sample size $n = 24$, the t-statistic is $t = 1.8$.

one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79



Two-sided p-value is between $2 \times 0.025 = 0.05$ and $2 \times 0.05 = 0.1$.

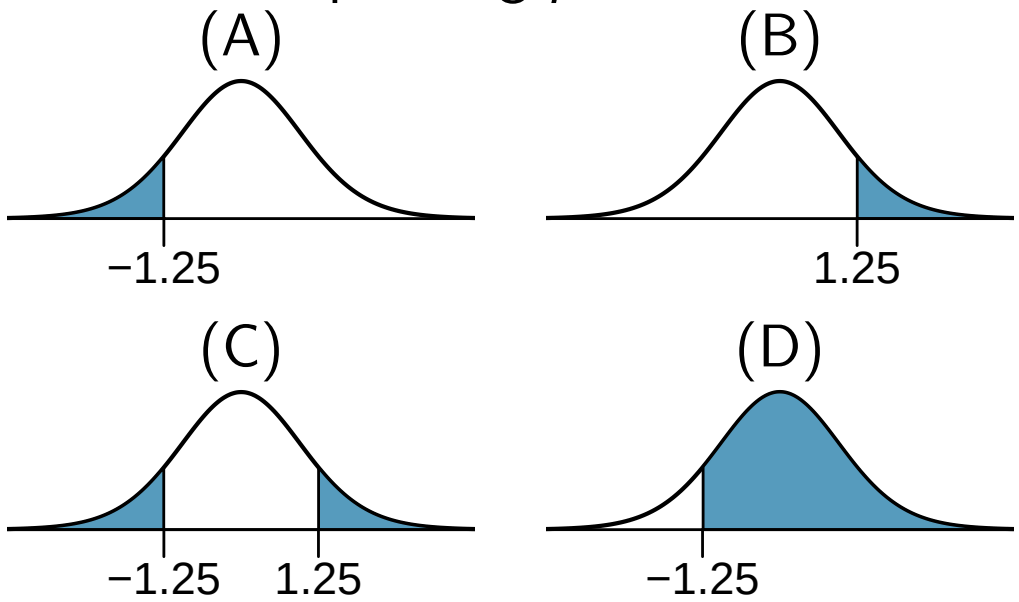
How to Use the t-Table to Find p-Values?

Example 3. Testing $H_0 : \mu = 20$ vs. $H_A : \mu > 20$, sample size $n = 57$, t-statistic $= 2.8$.

one tail		0.1	0.05	0.025	0.01	0.005	
two tails		0.2	0.10	0.050	0.02	0.010	
df	1	3.08	6.31	12.71	31.82	63.66	
	⋮	⋮	⋮	⋮	⋮	⋮	
	49	1.30	1.68	2.01	2.40	2.68	
	50	1.30	1.68	2.01	2.40	2.68	→
	60	1.30	1.67	2.00	2.39	2.66	→
	70	1.29	1.67	1.99	2.38	2.65	

How to Use the t-Table to Find p-Values?

Example 4. For testing $H_0 : \mu = 20$ vs. $H_A : \mu > 20$ with a sample of size $n = 17$ and a t-statistic $= -1.25$, which of the following areas represents the corresponding p-value?



T Test Example: Thermal Conductivity of Glass

Recall the example of *Thermal Conductivity of Glass*.

The thermal conductivity for 11 randomly selected pieces of a particular type of glass is measured.

`1.11, 1.07, 1.11, 1.07, 1.12, 1.08,`
`1.08, 1.18, 1.18, 1.18, 1.12`

We would like to test the claim "the mean conductivity of this type of glass is greater than 1."

Example: Thermal Conductivity of Glass — Hypotheses

- How shall we formulate the hypotheses?
- How to calculate the t-statistics?
- How to calculate the p-value? What conclusion can we make?

Example: Thermal Conductivity of Glass — p-value

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11

$$t_{obs} = 8.9 > 3.17 \quad \Rightarrow \quad \text{one sided } p\text{-value} < 0.005$$

Conclusion: We reject H_0 with p-value less than 0.005. The data provide convincing evidence that the mean conductivity of this type of glass is > 1.

It would be more interesting to know how big this difference is, which we can construct a confidence interval for it.

T-Tests and T-Confidence Intervals in R

```
> conduct = c(1.11,1.07,1.11,1.07,1.12,1.08,1.08,1.18,1.18,1.18,1.12)
> t.test(conduct, mu = 1, alternative = "greater")

        One Sample t-test

data:  conduct
t = 8.9538, df = 10, p-value = 2.167e-06
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 1.094259      Inf
sample estimates:
mean of x
 1.118182
```

Note that the 95% CI given (1.094259, Inf) $= (1.094259, \infty)$ is one-sided since we conducted a one-sided test.

T-Tests and T-Confidence Intervals in R

To conduct a "two-sided" test in R, change the "alternative" to "`two.sided`"

```
> t.test(conduct, mu = 1, alternative = "two.sided")

        One Sample t-test

data:  conduct
t = 8.9538, df = 10, p-value = 4.334e-06
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 1.088773 1.147591
sample estimates:
mean of x
 1.118182
```

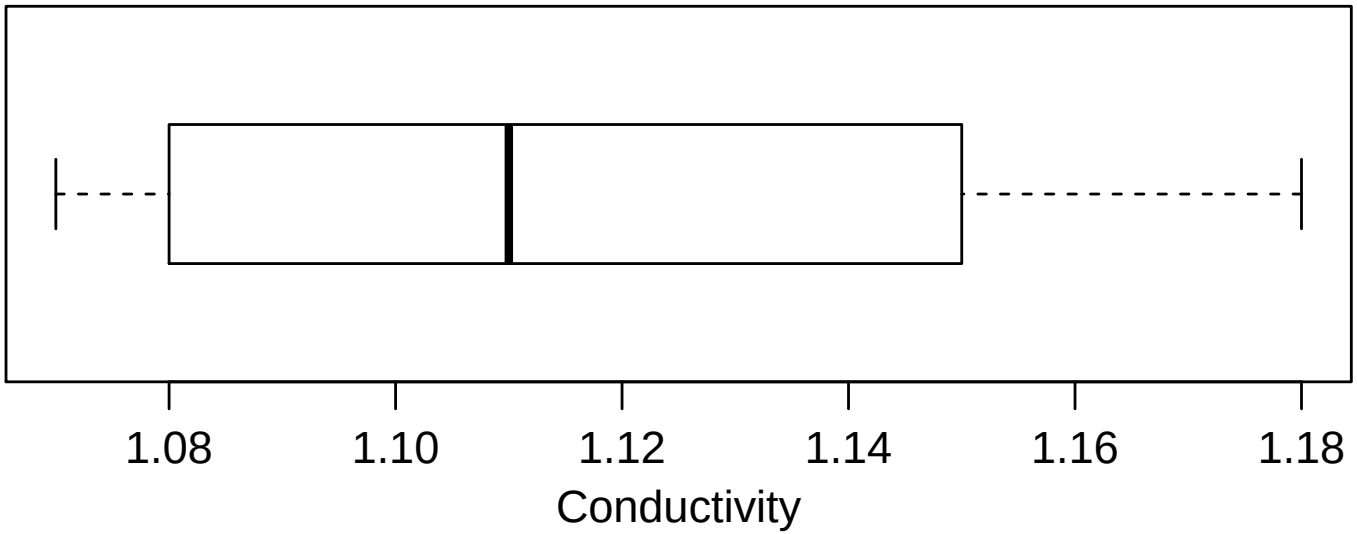
Conditions to Use t-Tests and t-Confidence Intervals

Though t-tests and t-confidence intervals don't require a big sample, they still require the following

- **Independence:** The observations should be independent.
- **Normality**:
  - For the t-statistic to have a t-distribution, the population distribution has to be normal, which is rarely true.
  - In particular, it's inherently difficult to verify normality in small data sets.
  - Fortunately, the t-test and t-CI have some **robustness against non-normality** (except in the case of outliers and strong skewness). Whereas, the impact diminishes as the sample size gets larger.

Checking Conditions for the Thermal Conductivity Example

- **Independence:** Suppose the observations are independent.
- **Normality**: The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not.



Example: Arsenic

- Arsenic is a toxic chemical element to humans and people can be exposed to it through contaminated drinking water, food, dust, and soil.
- Scientists have devised a non-invasive way to measure a person's level of arsenic poisoning: by examining toenail clippings.
- In a recent study, [1] data are reproduced from summary statistics and are approximate.

[1] M. Button, G. R. T. Jenkin, C. F. Harrington and M. J. Watts, "Human toenails as a biomarker of exposure to elevated environment arsenic," *Journal of Environmental Monitoring*, 2009; 11(3):610-617.

Example: Arsenic

- Scientists measured the level of arsenic (in mg/kg) in toenail clippings of $n = 8$ people who lived near a former arsenic mine in Great Britain as follows:
  0.8, 1.9, 2.7, 3.4, 3.9, 7.1, 11.9, 26.0.
- Suppose the 8 people examined were randomly sampled from residents near the former arsenic mine. Let μ be the the mean of Arsenic in toenail clippings for residents near the former arsenic mine.
- Is it legitimate to use t-test for $H_0 : \mu = 0$ v.s. $H_A : \mu > 0$?

Example: Arsenic

- Data Summary:

```
min  Q1 median  Q3 max   mean       sd n
0.8 2.5   3.65 8.3  26 7.2125 8.368041 8
```

- At such a small sample size ($n = 8$), a t-CI can be used only if **the population is fairly normal**.
- However, from the data summary we can see the sample is severely right-skewed (e.g., min/Q1 is much closer to the median than max/Q3 is), and there is an extreme outlier 26.0 that is over 3 IQRs above Q3.
- It's hence not legitimate to use a t-test.

Example: Utility Company Survey I

- A utility company serves 50,000 households. They would like to test the hypothesis that on average, each household has less than two TVs.
- As a part of a survey of customer attitudes, they take a SRS of 400 of these households.
- The average number of TVs in the sample households turns out to be 1.86, and the SD is 0.90.
- Let μ be the mean number of TVs in all 50,000 households.
- If possible, find a 95%-confidence interval for μ. If this isn't possible, explain why.

Example: Utility Company Survey I

- How would you do a hypothesis testing for this question?

Example: Utility Company Survey II

As part of the survey, all persons age 16 and over in the 400 sample households are interviewed. This makes 900 people.

On average, the sampled people watched 5.20 hours of TV the Sunday before the survey, and the SD was 4.50 hours.

True or False and explain: a 95%-confidence interval for the average number of hours spent watching TV by all persons age 16 and over in the 50,000 households on that Sunday is

$$\text{sample mean} \pm t^\star \times \frac{\text{sample SD}}{\sqrt{n}}$$

$$= 5.20 \pm 1.96 \times \frac{4.50}{\sqrt{900}} \approx 5.2 \pm 0.294$$

Example: Utility Company Survey II

- Population: all persons age 16 and over in the 50,000 households served by the utility company.
- Parameter: the average number of hours spent watching TV by all persons age 16 and over in the 50,000 households on that Sunday.
- The sample is NOT a **SRS** from the target population. There will be dependencies among the hours of TV watched among members of the same household.
- Hence we cannot construct the CI use the formula

$$\text{sample mean} \pm t^\star \times \frac{\text{sample SD}}{\sqrt{n}}$$

which assumes the observations in the sample were independent.

Relationship Between Confidence Intervals and Two-Sided Hypothesis Tests

Consider a two-sided test of the following hypotheses with the significance level α.

$$H_0 : \mu = \mu_0 \,, \quad H_a : \mu \neq \mu_0$$

- If μ_0 is a value inside the $(1 - \alpha)$ confidence interval for μ, then this test will have a p-value greater than α, and therefore will not reject H_0.
- If μ_0 is a value outside the $(1 - \alpha)$ confidence interval for μ, then this test will have a p-value smaller than α, and therefore will reject H_0.

Proof: Hypothesis tests and CI's

Connection

A level α two-sided test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value of μ_0 falls outside a $(1-\alpha)$ CI for μ.

Assume the test statistic is z and $2P(Z > |z|) = p < \alpha$. Let z_α be the critical value for level α. Assume the population SD is σ_0.

$$
\begin{aligned}
& p < \alpha \\
\iff & z > z_{\alpha/2} \quad \text{or} \quad z < -z_{\alpha/2} \\
\iff & \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2} \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \\
\iff & \mu_0 < \bar{x} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}} \quad \text{or} \quad \mu_0 > \bar{x} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}} \\
\iff & \mu_0 \notin [\bar{x} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}]
\end{aligned}
$$

μ_0 is not in the $1-\alpha$ confidence interval if and only if the null hypothesis is rejected at the α level.

Example

Suppose in a study,

- 90% CI for μ is $(4.81, 11.39)$;
- 95% CI for μ: $(4.18, 12.02)$;
- 99% CI for μ: $(2.95, 13.25)$.

Then

- $H_0 : \mu = 4$ is rejected at 5% level but not at 1% level
  (2-sided p-value is between 1% and 5%)
  because 4 is in the 99% CI but not in the 95% CI
- $H_0 : \mu = 4.5$ is rejected at 10% level but not at 5% level
  because 4.5 is in the 95% CI but not in the 90% CI

Summary: Hypothesis Testing for Population Mean

Step 1

Formulate the null hypothesis and the alternative hypothesis:

- The **null hypothesis** H_0 is the statement being tested. Usually it states that the difference between the observed value and the hypothesized value is only due to chance variation. For example, $\mu = 60$ minutes.
- The **alternative hypothesis** H_A is the statement we will favor if we find evidence that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values.
  For example, $\mu \neq 60$, $\mu > 60$, or $\mu < 60$.

A test is called

- **two-sided** if H_A is of the form $\mu \neq 60$.
- **one-sided** if H_A is of the form $\mu > 60$, or $\mu < 60$.

Step 2

Calculate the **test statistic** on which the test will be based.
The test statistic measures the difference between the observed data and what would be expected *if* the null hypothesis were true.

- When σ known, we use the Z-statistic,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- When σ unknown but n is large, we can replace σ with s in the Z-statistic.
- When σ unknown and n is small, we use the T-statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Step 3

Find the *p*-**value** of the observed data:

- The *p*-value is the probability of observing a test statistic *as extreme or more extreme than actually observed*, assuming the null hypothesis H_0 is true.
- The smaller the *p*-value, the stronger the evidence *against* the null hypothesis.
- if the *p*-value is as small or smaller than some number α (e.g. 0.01, 0.05), we say that the result is **statistically significant** at level α.
- α is called the **significance level** of the test.

How to calculate p-values

- In the Z-test, let $Z \sim N(0,1)$. The p-values for different alternative hypotheses:
  - $H_A : \mu > \mu_0$; p-value is $P(Z \geq z_{obs})$ (area of right-hand tail)
  - $H_A : \mu < \mu_0$; p-value is $P(Z \leq z_{obs})$ (area of left-hand tail)
  - $H_A : \mu \neq \mu_0$; p-value is $2P(Z \geq |z_{obs}|)$ (area of both tails)
- In the T-test, let $T \sim T_{n-1}$, the p-values for different alternative hypotheses:
  - $H_A : \mu > \mu_0$; p-value is $P(T \geq t_{obs})$ (area of right-hand tail)
  - $H_A : \mu < \mu_0$; p-value is $P(T \leq t_{obs})$ (area of left-hand tail)
  - $H_A : \mu \neq \mu_0$; p-value is $2P(T \geq |t_{obs}|)$ (area of both tails)

Test interpretations

- Saying that a result is statistically significant does not signify that it is large or necessarily important. That decision depends on the particulars of the problem.
- A statistically significant result only says that there is substantial evidence that H_0 is false.
- Failure to reject H_0 does not imply that H_0 is correct. It only implies that we have insufficient evidence to conclude that H_0 is incorrect.

Hypothesis Testing for Proportion

Statistical Inference for Proportion

Let $X \sim \text{Binom}(n, p)$.

- p is called the proportion parameter, typically unknown.
- How shall we estimate p?

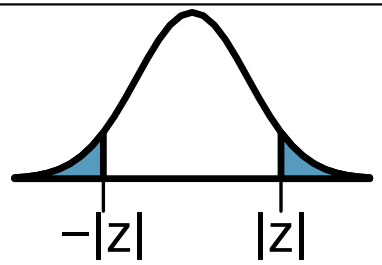
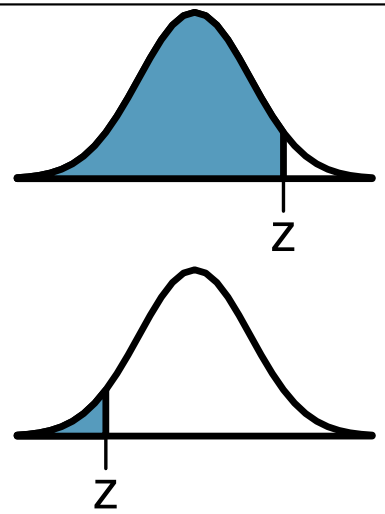
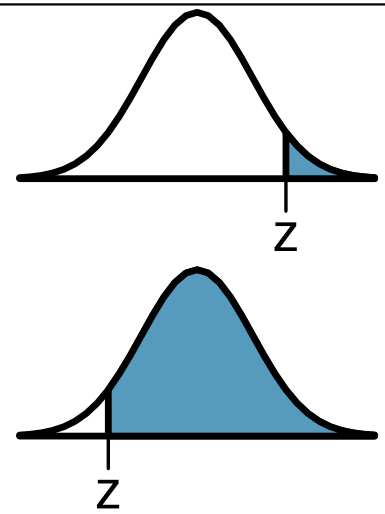
- Can we derive a $(1 - \alpha)$ level CI for p?

- Can we test $H_0 : p = p_0$ for some fixed value p_0?

Hypothesis Testing for Proportion (Large-Sample)

$H_0 : p = p_0$ for some fixed value p_0.

Under H_0, $Z = \dfrac{\widehat{p} - p_0}{SE} \dot{\sim} N(0,1)$, where $\text{SE} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$.

	Two-tailed test	One-tailed test	
H_a	$p \neq p_0$	$p < p_0$	$p > p_0$
p-value			

Here n should be so large that $np_0 \geq 10$, and $n(1-p_0) \geq 10$.

Remark. Recall that for confidence intervals, we use

$$SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

but for hypothesis testing we use

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Why?

- Recall by CLT when n is large $\widehat{p} \dot{\sim} N(p, p(1-p)/n)$
- When constructing CIs for p, p is unknown, so we estimate $p(1-p)/n$ by $\widehat{p}(1-\widehat{p})/n$
- Under $H_0 : p = p_0$, p is known to be p_0. There is no need to estimate p and the $p(1-p)/n$ is simply $p_0(1-p_0)/n$.

Example: Children's Play Preference

An observational study was conducted at Chicago Children's Museum to determine the age at which a child's preferred play partner switched from gender-neutral to a same-sex peer

- For 6-year old children, 59 of 97 preferred to interact with a same-sex peer (61%).

Under the null hypothesis of no preference, the probability that a child select a same-sex peer is $p = 0.5$.
We want to test if 6-year old children had a preference interacting with a same-sex peer.

Example: Children's Play Preference

We want to test $H_0 : p = 1/2$ versus $H_a : p > 1/2$.

- Is the Z test appropriate?
- Test statistic
- p-value
- Conclusion