STAT 135
7. Hypothesis testing

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30 (in person), Th 1:30-2:30 (virtual)
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

The null and alternative hypotheses

Hypothesis testing

Hypothesis testing is a method of using inference to test a hypothesis

Suppose the DMV claims that the average waiting time is 20 minutes. We want to test whether the average waiting time at the DMV is actually *more than 20 minutes.*

We want to test the null hypothesis:

$$H_0 : \mu = 20$$

Against the alternative hypothesis

$$H_1 : \mu > 20$$

We will use data from a random sample of waiting times and determine whether we have enough *evidence* to show that the average waiting time for the population is more than 20 minuets.

Hypothesis testing terminology

The terminology that we use when conducting a hypothesis test is that we either

a. Have enough evidence to reject the null hypothesis ($H_0 : \mu = 20$) in favor of the alternative hypothesis ($H_1 : \mu > 20$)

b. Don't have enough evidence to reject the null hypothesis ($H_0 : \mu = 20$)

Regardless, we are never ***proving*** that the null or alternative hypothesis is true.

The test statistic

Suppose that our data, $X_1, \ldots, X_n$, are IID from any distribution with variance σ^2

We want to test the null hypothesis: $H_0 : \mu = \mu_0$

Against the alternative hypothesis: $H_1 : \mu > \mu_0$

What statistic, $T(X_1, \ldots, X_n)$ could give us evidence to suggest whether we have evidence in favor of the alternative hypothesis?

The sample mean, $\bar{X}_n$!

But let's scale it:

The Z-**test statistic** is

$$Z = \frac{\bar{X}_n - \mu_0}{SD_{H_0}(\bar{X}_n)} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

Hypothesis testing example

If our hypothesis test is

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Under H_0:

$$Z = \frac{\text{estimated value – null value}}{\text{SD of estimate}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \underset{(approx)}{\overset{CLT}{\sim}} N(0,1)$$

If our test statistic looks unlikely to have come from a N(0, 1) distribution (e.g., because it's magnitude is very large), then this is evidence against H_0

Hypothesis testing example

Suppose that our data, $X_1, \ldots, X_n$, are IID from a distribution with $\sigma^2 = 4$

We want to test the null hypothesis: $H_0 : \mu = 5$

Against the alternative hypothesis: $H_1 : \mu > 5$

Suppose that from a sample of 20 data points, $X_1 = x_1, \ldots, X_{20} = x_{20}$, we observed $\bar{x}_n = 5.4$,

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{5.4 - 5}{2/\sqrt{20}} = 0.89$$

Is this big enough to claim we have evidence against H_0?

Probably not, our observed mean is less than 1 SD from the null mean

Hypothesis testing example

Suppose that our data, $X_1, \ldots, X_n$, are IID from a distribution with $\sigma^2 = 4$

We want to test the null hypothesis: $H_0 : \mu = 5$

Against the alternative hypothesis: $H_1 : \mu > 5$

Suppose that from a sample of 20 data points, $X_1 = x_1, \ldots, X_{20} = x_{20}$, we observed $\boxed{\bar{x}_n = \mathbf{7},}$

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{7 - 5}{2/\sqrt{20}} = 4.47$$

Is this big enough to claim we have evidence against H_0?

Probably!

Can we formalize this?

The p-value

The p-value

How do we determine what values of the test statistic *z* are “big” enough such that we can reasonably conclude that we have enough evidence against our null hypothesis?

P-value = the probability of observing a test statistic that is “as or more extreme” than z, *assuming the null hypothesis is true*

(where the definition of “extreme” is based on the alternative hypothesis)

The p-value is NOT the probability that the null is false,

Nor is it the probability that the alternative is true

P-value example

Suppose that our data, $X_1, \ldots, X_n$, are IID from a distribution with $\sigma^2 = 4$

We want to test the null hypothesis: $H_0 : \mu = 5$

Against the alternative hypothesis: $H_1 : \mu > 5$

Suppose that from a sample of 20 data points, $X_1, \ldots, X_{20}$, we observed $\bar{x}_n = 5.4$,

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{5.4 - 5}{2/\sqrt{20}} = 0.89$$

Under H_0:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \underset{approx}{\overset{CLT}{\sim}} N(0,1)$$

P-value = $P\left(Z \geq 0.89 \,|\, H_0\right) = 1 - \Phi(0.89) = 0.19$

Is this probability small enough to reject H_0?

Critical value and statistical significance

Critical value/significance level: The "critical value" or "significance level", α, is the value beyond which we reject the null hypothesis.

I.e., we reject the null hypothesis when the p-value is less than α

Convention says to reject the null hypothesis when the p-value is less than **0.05. We choose the significance level ourselves!**

i.e. the conventional significance level is $\alpha = 0.05$

Statistical significance: When the p-value is less than the significance level, α (e.g., p-value < 0.05), the result is said to be "statistically significant"

Computing the p-value in R

See z_test.R

Critical value and statistical significance

Assume that the null is *True* for the data you collected (I.e., the true population mean is μ_0).

What is the probability that your test (with significance level 0.05) will reject the null anyway?

5%

Rejection and acceptance regions

Rejection and acceptance regions:

The set of values of Z for which H_0 is rejected is called the **rejection region**

The set of values of Z for which H_0 is *not rejected* is called the **acceptance region**

Recall that we do not technically “accept” the null. We just gather evidence against it, and see if we have enough evidence to reject it.

Alternative hypothesis formats

Alternative hypotheses

There are several common forms of alternative hypotheses:

$$H_0 : \mu = \mu_0$$

One-sided tests

$$H_1 : \mu > \mu_0$$

$$H_1 : \mu < \mu_0$$

Two-sided test

$$H_1 : \mu \neq \mu_0$$

Composite hypotheses

$$H_1 : \mu = \mu_1$$

Simple hypothesis

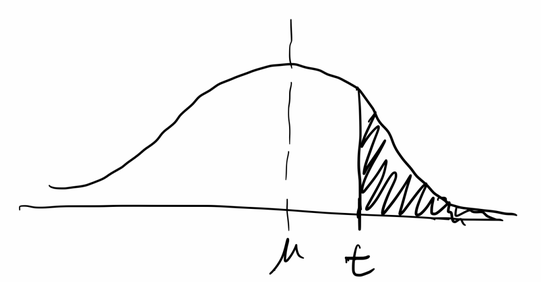
Alternative hypotheses

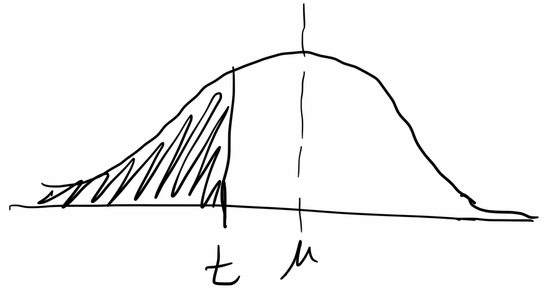
If the observed test statistic is $z = \dfrac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$

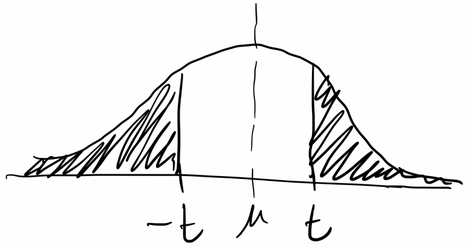
Under H_0:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \overset{CLT}{\sim} N(0,1)$$

Null	Alternative	**P-value**
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(Z \geq z \mid H_0) = 1 - \Phi(z)$
	$H_1 : \mu < \mu_0$	$P(Z \leq z \mid H_0) = \Phi(z)$
	$H_1 : \mu \neq \mu_0$	$P\left(\lvert Z \rvert \geq \lvert z \rvert \,\middle\vert\, H_0 \right) = 2(1 - \Phi(\lvert z \rvert))$







General useful result from the CLT

Theorem: if $X_1, X_2, \ldots, X_n$ is an IID sample from a population with mean μ and standard deviation σ, then:

$$P\left(|\overline{X} - \mu| \leq \delta\right) \approx 2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) - 1$$

Regardless of the original distribution of the X_i.

Proof:

$$P(|\overline{X} - \mu| \leq \delta) = P\left(\left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{\delta\sqrt{n}}{\sigma}\right) \overset{Z \sim N(0,1)}{\approx} P\left(|Z| \leq \frac{\delta\sqrt{n}}{\sigma}\right)$$

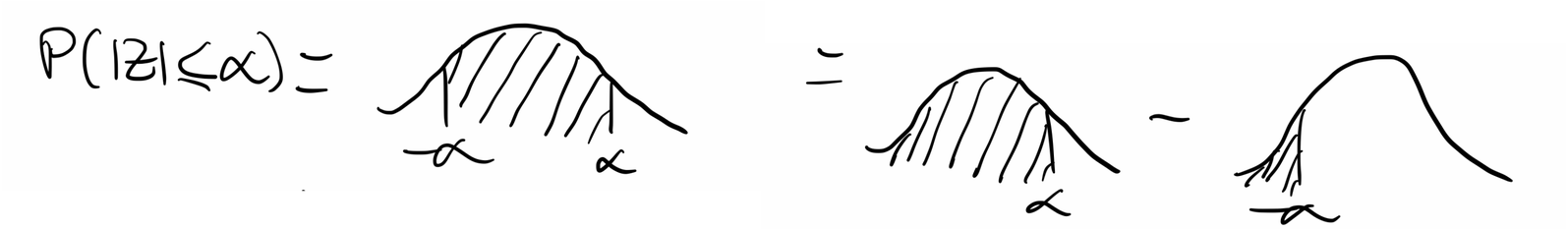
Divide both sides by $\sigma/\sqrt{n}$

$$= 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1$$

Since
$P(|Z| < \alpha) = 2\Phi(\alpha) - 1$

General useful result from the CLT

Visual proof that: $P(|Z| < \alpha) = 2\Phi(\alpha) - 1$



$= \Phi(\alpha) - [1 - \text{shaded}]$

$= \Phi(\alpha) - [1 - \text{shaded}]$

$= \Phi(\alpha) - [1 - \Phi(\alpha)]$

$= 2\Phi(\alpha) - 1$

Exercise

Corollary: if $X_1, X_2, \ldots, X_n$ is an IID sample from a population with mean μ and standard deviation σ, then:

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \geq \delta\right) \approx 2\left(1-\Phi\left(\delta\right)\right)$$

Regardless of the original distribution of the X_i.

Z-test example

Your friend claims that most people live within 15 minutes of campus. You disagree, you think that most people live closer to campus.

You randomly surveyed 40 people, and found that their average commute time was 13 minutes.

You happen to magically know that the true standard deviation of commute times is 3.5.

Do you have enough evidence to reject your friend's claim in favor of yours?

$n = 40 \quad \bar{x} = 13 \quad \sigma = 3.5$

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{13 - 15}{3.5/\sqrt{40}} = -3.6$$

$H_0 : \mu = 15$

$H_1 : \mu < 15$

P-value = $P\left(Z \leq -3.6 \,|\, H_0\right) = \Phi(-3.6) \approx 0.00016$

Which is less than 0.05, so reject H_0

Z-test example

Your friend claims that most people live within 15 minutes of campus. You disagree, you think that most people **don't live 15 mins** from

You randomly surveyed 40 people, and found that their average commute time was 13 minutes.

You happen to magically know that the true standard deviation of commute times is 3.5.

Do you have enough evidence to reject your friend's claim in favor of yours?

$n = 40 \quad \bar{x} = 13 \quad \sigma = 3.5$

$$Z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{13 - 15}{3.5/\sqrt{40}} = -3.6$$

$H_0 : \mu = 15$

$H_1 : \boldsymbol{\mu \neq 15}$ P-value = $P\left(|\boldsymbol{Z}| \geq 3.6 \,|\, H_0\right) = 2(1 - \Phi(3.6)) \approx 0.0003$

Which is less than 0.05, so reject H_0

Z-test examples

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \overset{CLT}{\underset{approx}{\sim}} N(0,1)$$

For a test with significance level $\alpha = 0.05$

$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Threshold p-value: $P\left(Z \geq 1.64\right) = 1 - \Phi(1.64) = 0.05$ Rejection region = $\{z \geq 1.64\}$ Acceptance region = $\{z < 1.64\}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Threshold p-value: $P\left(Z \leq -1.64\right) = \Phi(-1.64) = 0.05$ Rejection region = $\{z \leq -1.64\}$ Acceptance region = $\{z > -1.64\}$
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Threshold p-value: $P\left(\lvert Z\rvert \geq 1.96\right) = 2(1 - \Phi(1.96)) = 0.05$ Rejection region = $\{\lvert z\rvert \geq 1.96\}$ Acceptance region = $\{\lvert z\rvert < 1.96\} = \{-1.96 < z < 1.96\}$

Does this remind you of something?

Duality of hypothesis testing and confidence intervals

Confidence intervals for the sample mean

Confidence interval:

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$$

A 95% confidence interval for μ is:

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

P-value:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Acceptance region: $-1.96 \leq \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \leq 1.96$

Rearranged: $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$

If a 95% confidence interval for μ contains μ_0, then we would **not** reject H_0 (in favor of two-sided H_1) at the $\alpha = 0.05$ level

Confidence intervals and hypothesis tests

A $(1-\alpha)\%$ confidence interval is the set of values for which the null hypothesis of a two-sided test for the mean will not be rejected at level α

You randomly surveyed 40 people, and found that their average commute time was 13 minutes.

A 95% confidence interval for μ is: $\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right]$

$n = 40 \quad \bar{x} = 13 \quad \sigma = 3.5 \implies$ A 95% CI = $[11.9, 14.1]$

$H_0 : \mu = 15$

$H_1 : \boldsymbol{\mu \neq 15}$

This interval does not contain 15, so we would **reject** H_0 at the $\alpha = 0.05$ level

Type I and Type II errors

Type I and type II errors

Type I error: Rejecting the null hypothesis, H_0, when it is actually true

$\alpha =$ **Significance level/critical value:** (= probability of a type I error)

Type II error: Failing to reject the null hypothesis, H_0, when it is actually false

$\beta =$ Probability of a type II error

$1 - \beta =$ **Power:** Probability of rejecting the null when the null is False

Which error is more serious?

H_0 : Defendant innocent

H_1 : Defendant guilty

Which error is more serious?

H_0 : No cancer

H_1 : Cancer

Power

Sample size and power

Type II error: Failing to reject the null hypothesis, H_0, when it is actually false

P(Type II error) = β = P(do not reject $H_0 \mid H_0$ false)

Power = $1 - \beta$ = P(reject $H_0 \mid H_0$ false)

= P(reject $H_0 \mid H_1$ true)

Sample size and power

Power = $1 - \beta =$ P(reject $H_0 \,|\, H_1$ true)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

P-value = $P\left(Z > z \,|\, \boldsymbol{H_0}\right)$

Power = $P\left(Z > 1.64 \,|\, \boldsymbol{H_1}\right) = P\left(\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.64 \,\middle|\, \boldsymbol{H_1}\right)$

Is this N(0, 1) under H_1?

Sample size and power

Power = $1 - \beta =$ P(reject $H_0 \,|\, H_1$ true)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

Suppose that the true mean under H_1 is μ_1

$$\text{Power} = P\left(Z \geq 1.64 \,|\, H_1\right) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq 1.64 \,\middle|\, H_1\right)$$

$$= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \geq 1.64 \,\middle|\, H_1\right)$$

$$= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \geq 1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, H_1\right)$$

$$= 1 - \Phi\left(1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

Sample size and power

Power = $1 - \beta =$ P(reject $H_0 \,|\, H_1$ true)

Suppose that the true mean under H_1 is μ_1

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

Exercise: compute the power under different alternative hypotheses: $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$, and for different significance levels α

$$\text{Power} = P\left(Z \geq 1.64 \,|\, H_1\right) = 1 - \Phi\left(1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

- The power increases as n (sample size) increases
- The power increases as the difference between the mean under H_0 and H_1 $(\mu_1 - \mu_0)$ increases

Sample size power calculation

$$\text{Power} = P\left(Z \geq 1.64 \,|\, H_1\right) \;\; = 1 - \Phi\left(1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

What **sample size** do you need to have a **power of at least 0.9** to detect **differences between the null and alternative mean of at least 1**, assuming that $\sigma = 2$.

$$0.9 = 1 - \Phi\left(1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \implies q_{0.1} = 1.64 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

$$\implies \sqrt{n} = \frac{(1.64 - q_{0.1})\sigma}{\mu_1 - \mu_0}$$

$q_{0.1} = -1.28$

$$\implies n = (1.64 + 1.28)^2\sigma^2 \quad \overset{\sigma = 2}{\implies} \quad n \geq 34$$

set $\mu_1 - \mu_0 = 1$

T-test: Variance unknown, data normal

Variance unknown: t-test

Our original test statistic was

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \overset{CLT}{\underset{approx}{\sim}} N(0,1)$$

But we don't know σ!

Solution: use sample SD instead

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2}$$

If the X_i's are IID N(μ_0, σ^2), then

$$T = \frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

Note that this result assumes that our *data* is Normal

But in practice, even if our data is not normal… **it tends to approximately hold *for large sample sizes (at least n > 30)***

T-test example

Your friend claims that most people live within 15 minutes of campus. You disagree, you think that most people live further from campus.

You randomly surveyed 40 people, and found that their average commute time was 16.5 minutes.

The sample standard deviation is 3.5.

Do you have enough evidence to reject your friend's claim in favor of yours?

$n = 40 \quad \bar{x} = 16.5 \quad \hat{\sigma} = 3.5$

$$t = \frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{16.5 - 15}{3.5/\sqrt{40}} = 2.7$$

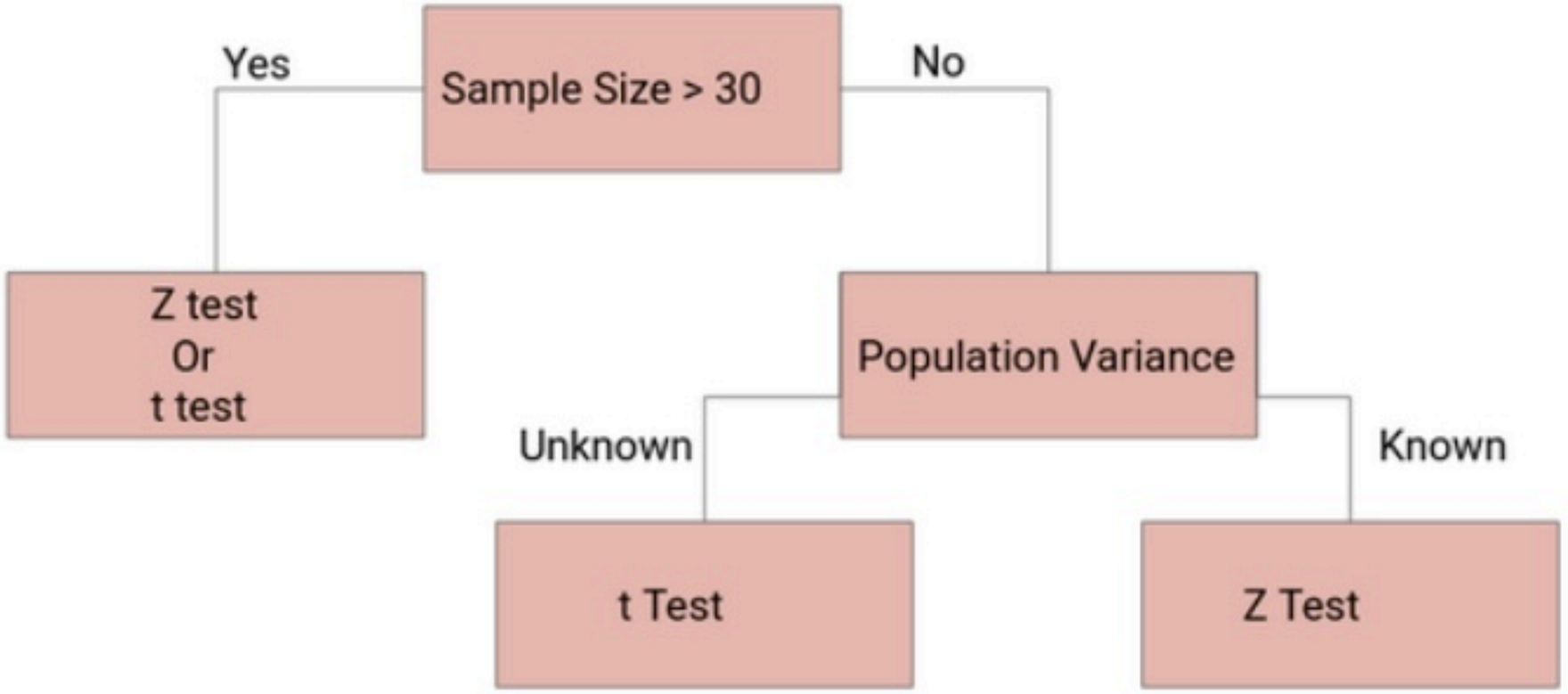
$H_0 : \mu = 15$

$H_1 : \mu > 15$

t_{39} P-value $= P\left(T \geq 2.7 \,|\, H_0\right) \approx 0.005$

Normal P-value $= P\left(T \geq 2.7 \,|\, H_0\right) = 1 - \Phi(2.7) \approx 0.003$

T-test vs Z-test in practice



Source: https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/

T-test in R

See t_test.R

Test for proportions

Test for proportions: normal approximation

A popular weight-loss program claims that 50% of their clients lose 10lb in the first two weeks of the program.

You are skeptical, and you want to test this theory. You collect some data on a random sample of 40 program participants and record whether they lost 10lb in the first two weeks or not. You found that 16 of them did.

Do you have enough evidence to reject the claim that 50% of the clients lose 10lb in the first two weeks of the program?

$$H_0 : p = 0.5 \qquad n = 40$$

$$H_1 : p \neq 0.5 \qquad \hat{p} = 16/40 = 0.4$$

Note the sample size is > 30 so we will use a normal approximation

$$z = \frac{\text{estimate} - \text{null value}}{SD_{H_0}(\text{estimate})} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.4 - 0.5}{\sqrt{0.5 \times 0.5/40}} = -1.26$$

$$\text{P-value} = P(|Z| \geq 1.26 \,|\, H_0) = 2(1 - \Phi(1.26)) = 0.21$$

Test for proportions in R

See prop_test.R

Testing for normality: Q-Q plots

Discussion: when is normality assumed?

Is normality assumed for a Z-test of the mean (known variance)?

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \overset{CLT}{\sim} N(0,1)$$

No, the CLT says that the sample mean is normally distributed, regardless of the original distribution (as long as the data are IID)

Is normality assumed for a t-test of the mean (unknown variance)?

$$T = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{df}$$

Yes, in order for the test statistic to have a t-distribution, the data needs to be normal.

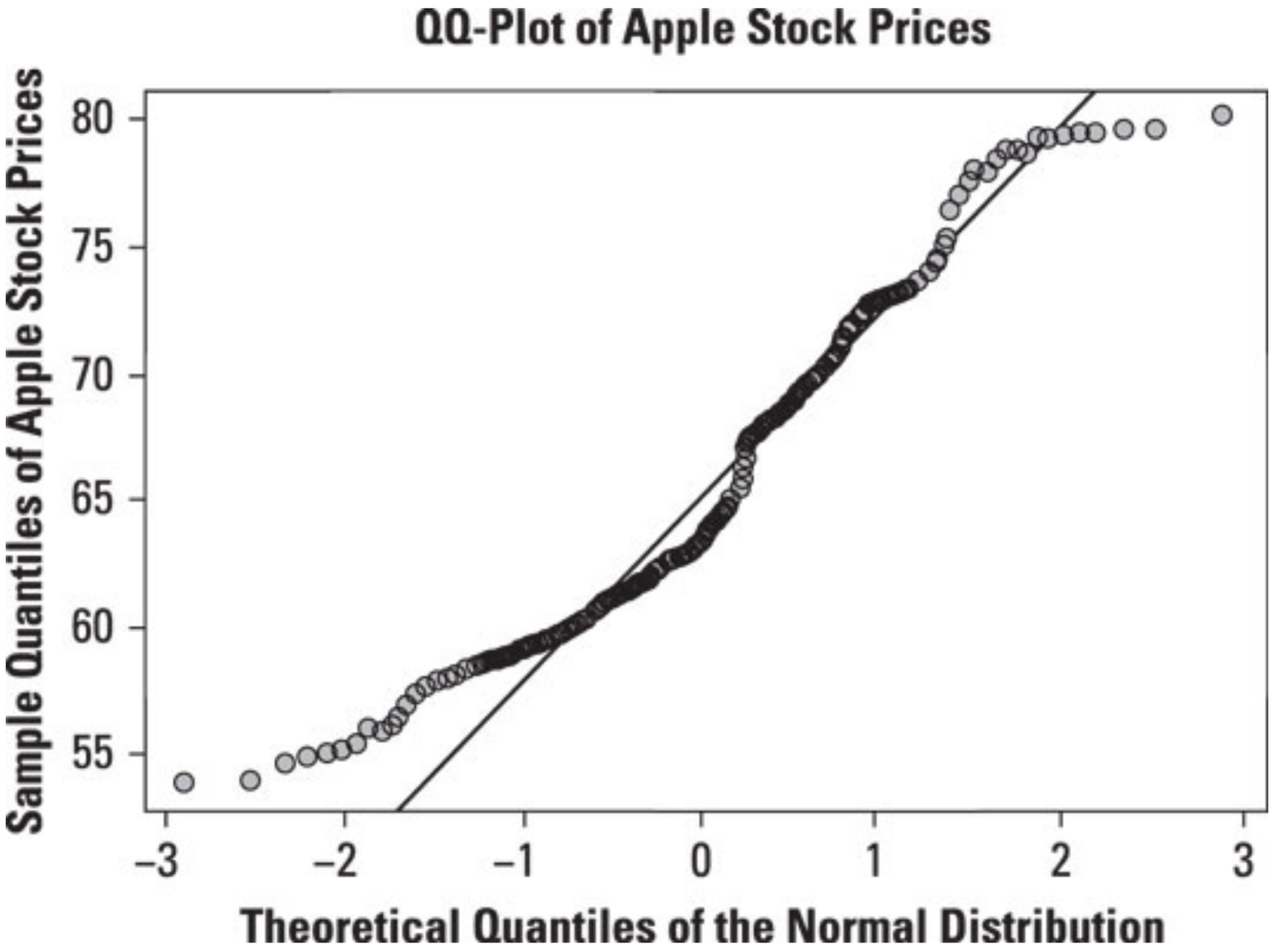
But if the data is non-normal, the statistic has an approximate t-distribution when there are at least 30 or so data points.

QQ-plot

To test whether some data comes from a particular distribution…

A QQ-plot plots the data's quantiles against the theoretical quantiles of the distribution of interest.

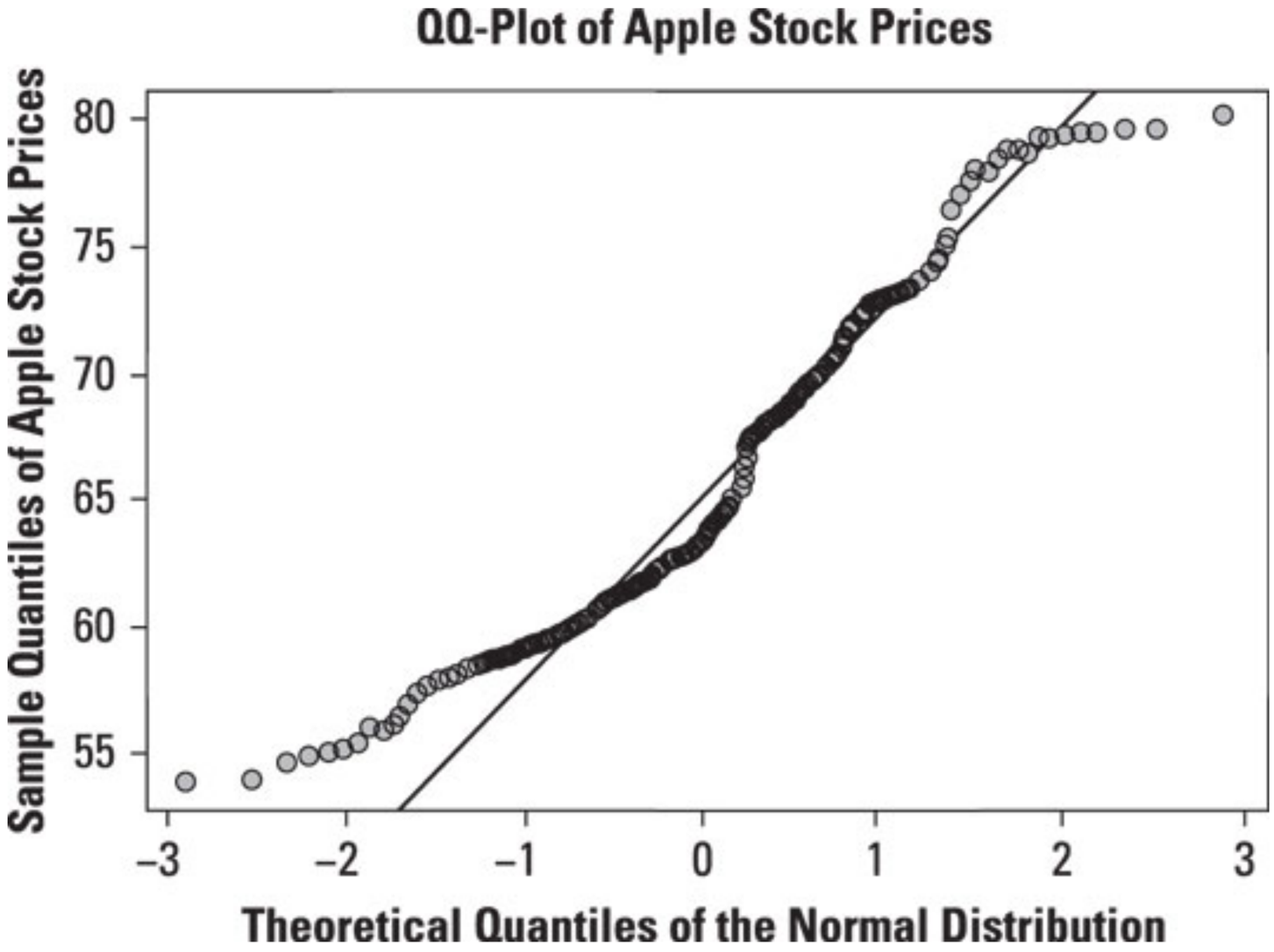
If the points follow a diagonal straight line, then they are probably pretty similar to the distribution



QQ-plot

These points do not follow the straight line at all…

Apple stock prices are probably not normal



QQ-plot R example

See qq.R

Non-parametric hypothesis test

Non-parametric hypothesis tests

There exist several non-parametric versions of these hypothesis tests that do not assume the data has a Normal distribution (although the t-test is ok for non-normal data when the sample size is large enough).

These tests consider the *rankings* of the observations

E.g. the **Wilcoxon signed-rank test** (https://en.wikipedia.org/wiki/Wilcoxon_signed-

In practice, two-sample tests are a lot more common than one sample tests, so we will introduce non-parametric tests in the context of two-sample tests, rather than one-sample tests

Non-parametric tests are often **less powerful** than their parametric counterpart