6.3 Hypothesis testing

Last section illustrated how probability can enable confidence interval estimation. We can also use probability as a means to use data to quantitatively assess the plausibility of a trial value of a parameter.

Statistical inference is using data from the sample to draw conclusions about the population. $(\bar{X} \xrightarrow{\text{infer}} \mu)$

1. Interval estimation (confidence intervals)
   estimating a population parameter and specifying the degree of precision of the estimate
   • what is μ? $\bar{X} \rightarrow \mu \in [3,5]$
2. Hypothesis testing
   testing the validity of statements about the population parameter
   • Is $\mu > 4$? use $\bar{x} \rightarrow$ "yes" $\mu > 4$, or "no" μ not > 4.

Definition 6.3. Statistical *significance testing* is the use of data in th quantitative assessment of the plausibility of some trial value for a parameter (or function of one or more parameters).

i.e. assess plausibility of a process mean value of 138.9 for fill weight of baby food

Significance (or hypothesis) testing begins with the specification of a trial value (or **hypothesis**).

Definition 6.4. A *null hypothesis* is a statement of the form

$$\underbrace{\text{Parameter}}_{\text{(not statistic)}} = \#$$

or

$$\text{Function of parameters} = \#$$

for some $\overset{\text{constant}}{\#}$ that forms the basis of investigation in a significance test. A null hypothesis is usually formed to embody a status quo/"pre-data" view of the parameter. It is denoted H_0.

"null" because it is a statement of no difference (equality).

Definition 6.5. An *alternative hypothesis* is a statement that stands in opposition to the null hypothesis. It specifies what forms of departure from the null hypothesis are of concern. An alternative hypothesis is denoted as H_a. It is of the form

$$\text{Parameter} \neq \# \qquad \text{or} \qquad \text{Parameter} > \# \qquad \text{or} \qquad \text{Parameter} < \#$$

Examples (testing the true mean value):

1	2	3
$H_0 : \mu = \#$	$H_0 : \mu = \#$	$H_0 : \mu = \#$
$H_a : \mu \neq \#$	$H_a : \mu > \#$	$H_a : \mu < \#$
two-sided	one-sided	

Often, the alternative hypothesis is based on an investigator's suspicions and/or hopes about th true state of affairs.

The **goal** is to use the data to debunk the null hypothesis in favor of the alternative.

1. Assume H_0. ("status quo")
2. Try to show that, under H_0, the data are "preposterous." (use probability)
3. If the data are preposterous, reject H_0 and conclude H_a.

The outcomes of a hypothesis test consists of:

	Decision: H_0	Decision: H_A
Truth: H_0	ok	Type I error
Truth: H_A	Type II error	ok

same as α from CI's (probability of rejecting H_0 when H_0 is actually true)

specify α before we do significance testing

P(H_0 true, but you reject it) = α

Type I error probability is fixed before you do testing (before you look at data)

Example 6.11 (Fair coin). Suppose we toss a coin $n = 25$ times, and the results are denoted by $X_1, X_2, \ldots, X_{25}$. We use 1 to denote the result of a head and 0 to denote the results of a tail. Then $X_1 \sim Binomial(1, \rho)$ where ρ denotes the chance of getting heads, so $\mathrm{E}(X_1) = \rho, \mathrm{Var}(X_1) = \rho(1-\rho)$. Given the result is you got all heads, do you think the coin is fair?

Null hypothesis: $H_0: p = 0.5$ "Coin is fair"

$H_A: p \neq 0.5$ "Coin is not fair"

Binomial formula (pmf)

If H_0 is correct, $P(\text{result is all heads}) = \left(\frac{1}{2}\right)^{25} < 0.000001$

$\Rightarrow$ I don't think this coin is fair (reject H_0)

In the real life, we may have data from many different kinds of distributions! Thus we need a universal framework to deal with these kinds of problems.

We have $n = 25 \geq 25$ independent and identically distributed trials

$\Rightarrow$ by the CLT if $H_0: \rho = 0.5$, then:

0.5 if H_0 true.

$\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$ where $X_i \sim Binom(1, \rho)$

$EX_i = \rho \Rightarrow E\bar{X} = \rho$

$Var X_i = \rho(1-\rho) \Rightarrow Var\bar{X} = \rho(1-\rho)/25$

by CLT

$\bar{X} \dot{\sim} N\left(\rho, \frac{\rho(1-\rho)}{25}\right)$

$\frac{\bar{X} - \rho}{\sqrt{\rho(1-\rho)/25}} \dot{\sim} N(0,1)$.

We observe $\bar{x} = 1$. So,

$\frac{\bar{x} - 0.5}{\sqrt{0.5(1-0.5)/25}} = 5$ Then $P(Z \text{ bigger than } 5 \text{ or less than } -5) < .000001$!

This is absurd, so we reject H_0.

6.3.1 Significance tests for a mean

Definition 6.6. A *test statistic* is the particular form of numerical data summarization used in a significance test.

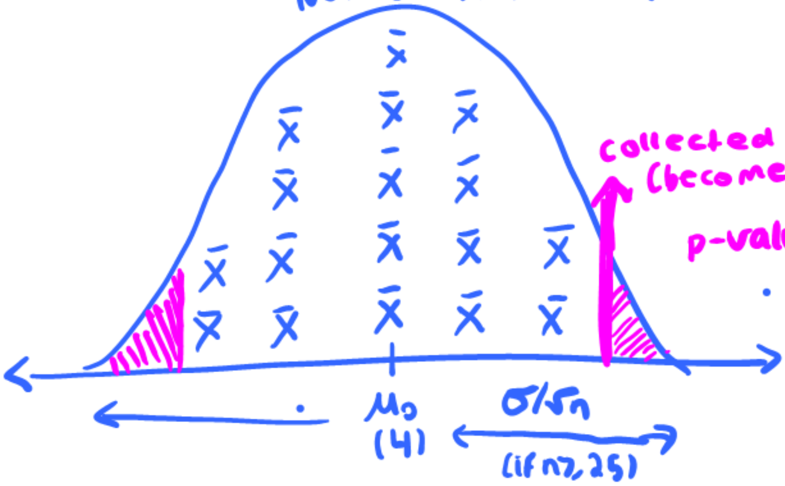
Definition 6.7. A *reference (or null) distribution* for a test statistic is the probability distribution describing the test statistic, provided the null hypothesis is in fact true.

If sample means and $n \geq 25 \longrightarrow N(0,1)$

Definition 6.8. The *observed level of significance or p-value* in a significance test is the probability that the reference distribution assigns to the set of possible values of the test statistic that are at least as extreme as the one actually observed.

(in the previous problem, the p-value was < 0.000001)

Null distribution (comes from assuming Ho is true, $H_0: \mu = \mu_0$) ($\mu = 4$)



Let k be test statistic

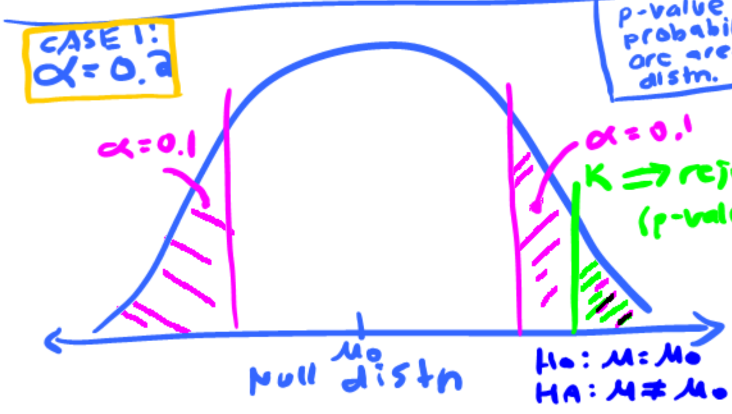
Say $H_0: \mu = \mu_0 = 4$

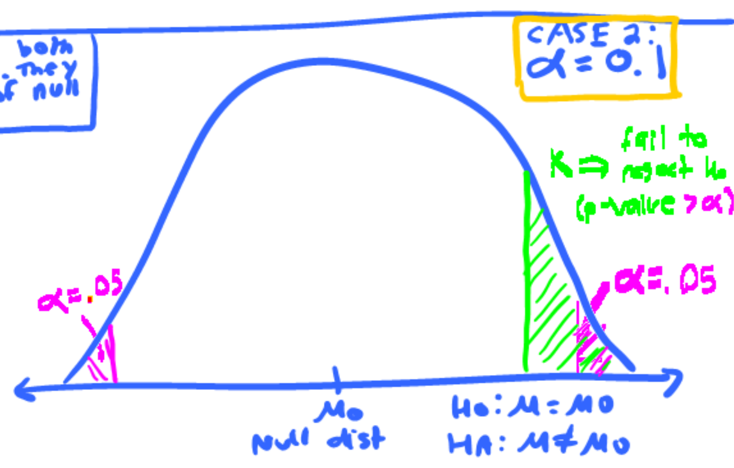
$H_A: \mu \neq \mu_0 \neq 4$

p-value = P(see data "as extreme as" k if Ho is true)

$= P(Z < -k \text{ or } Z > k)$

p-value and α are both probabilities $[0,1]$. They are areas in tails of null distn.





Based on our results from Section 6.2 of the notes, we can develop hypothesis tests for the true mean value of a distribution in various situations, given an iid sample $X_1, \dots, X_n$ where $H_0 : \mu = \mu_0$.

Let K be the value of the test statistic, $Z \sim N(0,1)$, and $T \sim t_{n-1}$. Here is a table of p-values that you should use for each set of conditions and choice of H_a.

	Situation	K	$H_a : \mu \neq \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	
use CLT	$n \geq 25, \sigma$ known	$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$	$P(\lvert Z\rvert > K)$	$P(Z < K)$	$P(Z > K)$	reference dsn is Normal
	$n \geq 25, \sigma$ unknown	$\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$	$P(\lvert Z\rvert > K)$	$P(Z < K)$	$P(Z > K)$	
	$n < 25, \sigma$ unknown	$\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$	$P(\lvert T\rvert > K)$	$P(T < K)$	$P(T > K)$	reference dsn is t_{n-1}

Steps to perform a hypothesis test:

1. State H_0 and H_A
2. State α, significance-level (usually 0.1, 0.05, 0.01)
3. State the form of test-statistic, distn under the null hypothesis, and assumptions
4. Calculate the test statistic and p-value
5. Make a decision based on the p-value.
   - if p-value is $< \alpha \Rightarrow$ reject H_0 otherwise fail to reject H_0
6. Interpret the conclusion using layman's terms
   rejected $H_0 \Rightarrow$ have evidence for H_a (in context)
   fail to reject $H_0 \Rightarrow$ do not have evidence for H_a (in context).

Example 6.12 (Cylinders). The strengths of 40 steel cylinders were measured in MPa. The sample mean strength is 1.2 MPa with a sample standard deviation of 0.5 MPa. At significance level $\alpha = 0.01$, conduct a hypothesis test to determine if the cylinders meet the strength requirement of 0.8 MPa.

1. $H_0: \mu = 0.8$
   $H_A: \mu > 0.8$

2. $\alpha = 0.01$

3. Since σ is unknown, and $n = 40 \geq 25$,
   $$K = \frac{\bar{x} - 0.8}{S/\sqrt{n}}$$ is the test statistic.

   I assume $X_1, \ldots, X_{40}$ are iid with mean μ and variance σ^2.
   Then by CLT, $K \dot{\sim} N(0,1)$ under the null hypothesis.

4. $$K = \frac{1.2 - 0.8}{0.5/\sqrt{40}} = 5.06$$

   p-value. $$P(Z > 5.06) = 1 - P(Z \leq 5.06) = 1 - \Phi(5.06) \approx 1 - 1 = 0$$

5. Since p-value $<< \alpha$, I reject H_0 in favor of H_A.

6. There is overwhelming evidence to conclude that the cylinders meet the strength requirement of 0.8 MPa.

Example 6.13 (Concrete beams)**.** 10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

The sample mean was 9.2 MPa and the sample variance was 3.0933 MPa. Conduct a hypothesis test to find out if the flexural strength is different from 9.0 MPa.

1. $H_0: \mu = 9.0$, $H_A: \mu \neq 9.0$

2. Choose $\alpha = 0.05$

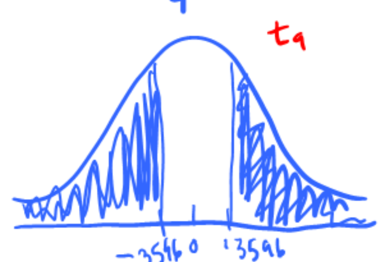
3. I will choose test statistic $K = \dfrac{\bar{x} - 9.0}{s/\sqrt{n}}$ (unknown σ)

Since $n = 10 < 25$, (small), we must assume $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Then if our assumptions hold, $K \sim t_{n-1} = t_9$ under H_0.

4. $K = \dfrac{9.2 - 9}{\sqrt{\frac{3.0933}{10}}} = 0.3596$

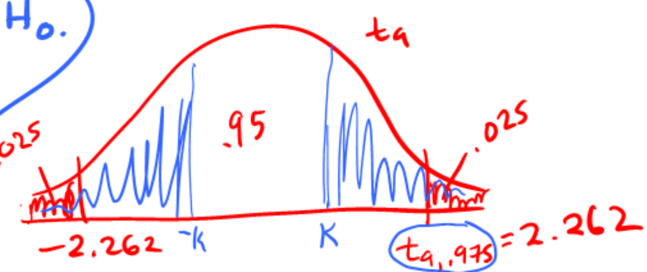
p-value $P(|T| > 0.3596)$ where $T \sim t_9$



Since $K = 0.3596 < t_{9,.975} = 2.262$,

we know $P(|T| > .3596) > .05$

5. Since p-value $> .05$, I fail to reject H_0.



6. There is not enough evidence to conclude the true mean flexural strength of the beams is different from 9.0 MPa.

6.3.2 Hypothesis testing using the CI

We can also use the $1-\alpha$ confidence interval to perform hypothesis tests (instead of p-values). The confidence interval will contain μ_0 when there is little to no evidence against H_0 and will not contain μ_0 when there is strong evidence against H_0.

Steps to perform a hypothesis test using a confidence interval:

1. State the hypotheses H_0 and H_A
2. State the significance level, α
3. State the form of the $1-\alpha$ CI along with all assumptions
   - use a one-sided CI for 1-sided tests (i.e. $H_A: \mu < \#$ or $H_A: \mu > \#$)
   - use a two-sided CI for two-sided tests (i.e. $H_A: \mu \neq \#$).
4. Calculate the $1-\alpha$ CI
5. (make a decision) Based on $1-\alpha$ CI, either reject H_0 or fail to reject H_0
6. Interpret the conclusion using layman's terms in the context of the problem.

Example 6.14 (Breaking strength of wire, cont'd)**.** Suppose you are a manufacturer of construction equipment. You make 0.0125 inch wire rope and need to determine how much weight it can hold before breaking so that you can label it clearly. You have breaking strengths, in kg, for 41 sample wires with sample mean breaking strength 91.85 kg and sample standard deviation 17.6 kg. Using the appropriate 95% confidence interval, conduct a hypothesis test to find out if the true mean breaking strength is above 85 kg.

① $H_0: \mu = 85$, $H_A: \mu > 85$ μ is true mean breaking strength.

② $\alpha = 0.05$

③ One sided test, where we care about the lower bound. I will use the $1-\alpha$ CI

$$\left(\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}, \infty\right)$$

because $n = 41 \geq 25$ and σ^2 unknown.

I am assuming the data points are iid draws from a dsn w/ mean μ and variance σ^2.

Since $n \geq 25$, we use the $z_{1-\alpha}$ quantile.

④ from example 6.7, the CI is $(87.3422, \infty)$ $\Rightarrow$ reject H_0. $\mu_0 = 85$ is not in this interval

⑤ With 95% confidence we have shown the $\mu > 87.3422$. Hence at significance level $\alpha = .05$, we can reject H_0 in favor of H_A.

⑥ There is significant evidence to show that the true mean breaking strength of the wire is greater than 85 kg. The requirement seems to be met.

Example 6.15 (Concrete beams, cont'd)**.** 10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

The sample mean was 9.2 MPa and the sample variance was 3.0933 MPa. At $\alpha = 0.01$, test the hypothesis that the true mean flexural strength is 10 MPa using a confidence interval.

Example 6.16 (Paint thickness, cont'd)**.** Consider the following sample of observations on coating thickness for low-viscosity paint.

[1] 0.83 0.88 0.88 1.04 1.09 1.12 1.29 1.31 1.48 1.49 1.59 1.62 1.65 1.71 [15] 1.76 1.83

Using $\alpha = 0.1$, test the hypothesis that the true mean paint thickness is 1.00 mm. Note, the 90% confidence interval for the true mean paint thickness was calculated from before as $(1.201, 1.499)$.