Section 6.2 - Tests of Significance

Statistics 104

Autumn 2004

Tests of Significance

Example: VCR tapes

Stated mean time: 360 minutes

Standard deviation: 8 minutes

The confidence interval analysis earlier suggested a problem. Suppose that they thought that they fixed the problems and wanted to do another study. As before, they sampled 64 tapes and observed $\bar{x} = 358.5$ minutes.

This looks better.

We now want to look at the question of when a sample mean like this is consistent or inconsistent with the desired mean level.

Lets assume that the true mean level is really 360 minutes.

What is the probability of getting a sample average at least as far from 360 minutes as the observed $\bar{x} = 358.5$.

Answer: the probability of getting a $\bar{x}$ further away from 360 than the observed $\bar{x} = 358.5$ is 0.1336. (Justification to come)

We are interested if the data is consistent with a particular mean value.

Null Hypothesis:

> The statement being tested in a significance test. The test of significance is designed to assess the strength of evidence against the null hypothesis.
>
> Usually the null hypothesis is a statement of "no effect" or "no difference".

The null hypothesis is usually abbreviated by H_0.

> $H_0 : \mu = 360$ minutes.

Alternative Hypothesis:

The statement we hope or suspect is true instead of H_0. It describes the deviations from H_0 of interest

The null hypothesis is usually abbreviated by H_0.

H_A : $\mu < 360$ minutes (worried about fraudulent claims)

The alternative hypothesis can either be one-sided, like above, or two-sided. We could also look at

H_A : $\mu \neq 360$ minutes

This would be appropriate when you are interested in both of

$\mu < 360$ (fraud) and $\mu > 360$ (wasteful)

Example: Cheese making

A cheese maker was worried that milk from one of his suppliers was being watered down.

Pure milk: mean freezing temperature = -0.545°C

Watered milk: mean freezing temperature > -0.545°C

Adding water won't decrease the freezing point.

Hypotheses to be tested:

$$H_0 : \mu = -0.545$$

$$H_A : \mu > -0.545$$

The one-sided alternative makes more sense here.

However usually you want to look at two-sided alternatives.

Example: Testing blood pressure medicine

H_0: mean change $= 0$

H_A: mean change $\neq 0$

Interested in both here as (depending on how things were measured)

Mean change < 0 (beneficial)

Mean change > 0 (harmful)

H_0 and H_A are **ALWAYS** in terms of population parameters, not sample statistics

$H_0 : \bar{x} = 0$	Wrong
$H_A : \bar{x} \neq 0$	Don't Do

Test Statistic:

The statistic that we want to use to examine the 2 hypotheses. Usually this will be the estimate of the parameter of interest (or based on it).

Values of the estimate far away form the parameter value specified by H_0 give evidence against the null hypothesis. The alternative hypothesis determines which values count against the null.

We will use $\bar{x}$ to examine hypotheses about the population mean μ.

Question:

In the VCR example, what's the probability of getting a sample average more extreme than observed, which in this case is less than or equal to 358.5 minutes. (Averages that are more consistent with fraudulent claims)

Answer:

Standard error of the sampling distribution

$$SE = \frac{8}{\sqrt{64}} = 1$$

$$z = \frac{358.5 - 360}{1} = -1.5$$

$$P[\bar{X} \leq 358.5] = P[Z \leq -1.5] = 0.0668$$

What is more extreme?

It depends on the alternative hypothesis

For a one-sided hypothesis

$$H_A : \mu > \mu_0 \qquad \bar{x} \geq \bar{x}_{obs}$$
$$H_A : \mu < \mu_0 \qquad \bar{x} \leq \bar{x}_{obs}$$

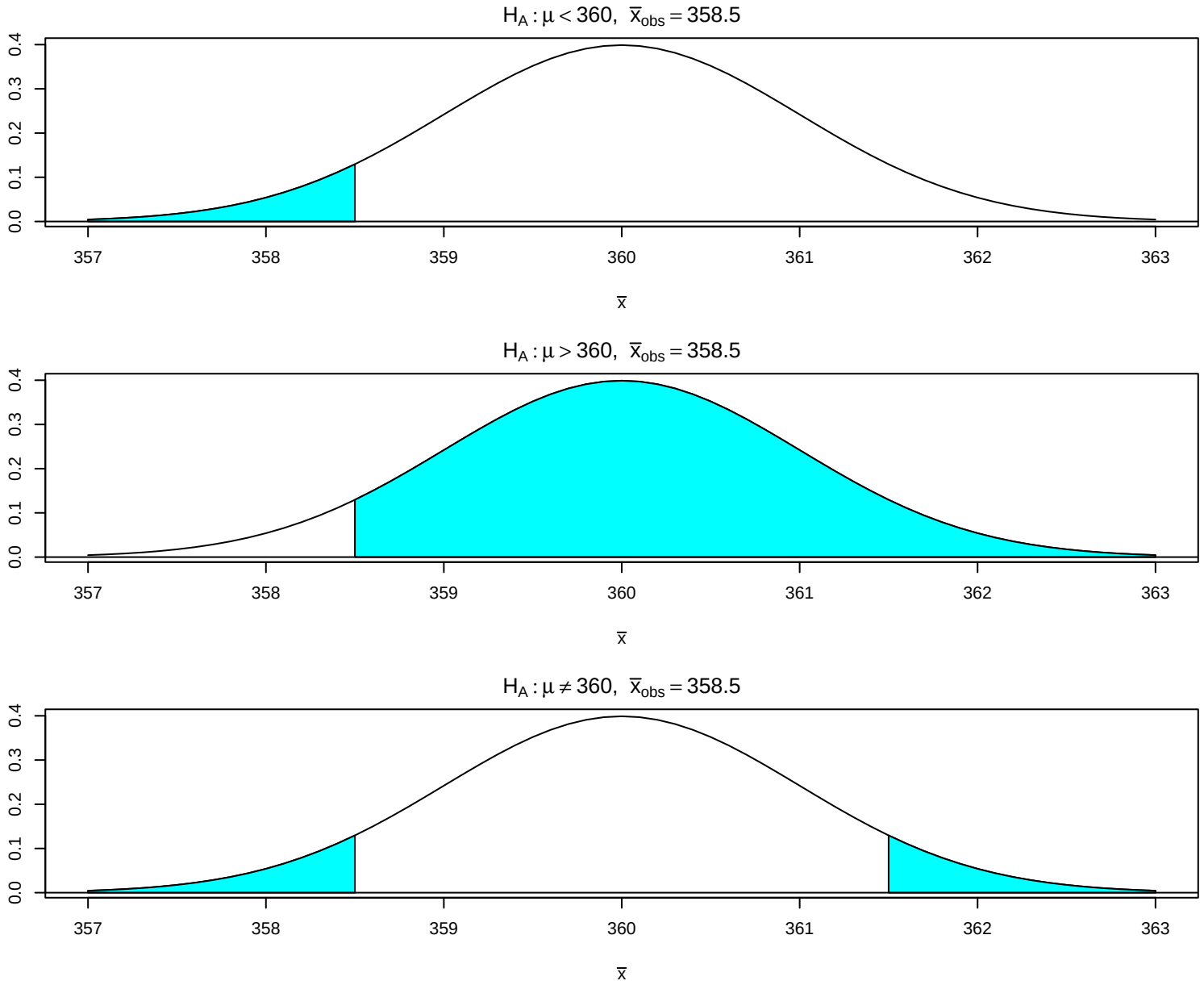
For a two-sided alternative ($H_A : \mu \neq \mu_0$) it's a bit more complicated. Suppose that $\delta = |\mu_0 - \bar{x}_{obs}|$. Then more extreme is

$$\bar{x} \leq \mu_0 - \delta \quad \text{or} \quad \bar{x} \geq \mu_0 + \delta$$

For the VCR example with an alternative hypothesis of $H_A : \mu \neq 360$, more extreme is

$$\bar{x} \leq 358.5 \quad \text{or} \quad \bar{x} \geq 361.5$$

(more than 1.5 minutes from the null hypothesis value)


$H_A : \mu < 360,\ \bar{x}_{obs} = 358.5$
$H_A : \mu > 360,\ \bar{x}_{obs} = 358.5$
$H_A : \mu \neq 360,\ \bar{x}_{obs} = 358.5$
0.0
0.1
0.2
0.3
0.4
357
358
359
360
361
362
363
$\bar{x}$

The probabilities of falling into the more extreme regions, assuming that H_0 is true are of interest.

$$P[\bar{X} \leq 358.5] = P[Z \leq -1.5] = 0.0668$$

$$\begin{aligned} P[\bar{X} \leq 358.5 \text{ or } \bar{X} \geq 361.5] &= P[Z \leq -1.5] + P[Z \geq 1.5] \\ &= 2P[Z \leq -1.5] \\ &= 0.1336 \end{aligned}$$

P-value:

The probability, assuming that H_0 is true, that the test statistic would take a value as or more extreme than that actually observed. The smaller the p-value, the stronger the evidence against H_0 provided by the data. The alternative hypothesis indicates which values are considered as or more extreme.

So for the tape example the p-value $= 0.0668$ for the one-sided fraud alternative and p-value $= 0.1336$ for the two-sided alternative.

In this situation, an easier test statistic to use over the sample mean $\bar{x}$, is its standardized value

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Then more extreme for the three situations are

$$\begin{array}{ll} H_A : \mu > \mu_0 & z \geq z_{obs} \\ H_A : \mu < \mu_0 & z \leq z_{obs} \\ H_A : \mu \neq \mu_0 & z \geq |z_{obs}| \text{ or } z \leq -|z_{obs}| \end{array}$$

Based on this, the p-values for the z-test are

$$\begin{array}{ll} H_A : \mu > \mu_0 & P[z \geq z_{obs}] \\ H_A : \mu < \mu_0 & P[z \leq z_{obs}] \\ H_A : \mu \neq \mu_0 & 2P[z \geq |z_{obs}|] \end{array}$$

What is a small p-value?

As mentioned earlier, the smaller the p-value, the stronger the evidence against H_0 provided by the data. What do we want to consider as a small p-value?

Statistical Significance:

> If the p-value is as small or smaller than α, we say that the data are **statistically significant at level α**
>
> α is known as the significance level.

Popular choices for α are 0.05 and 0.01.

Picking an α for a hypothesis test is equivalent to picking a C for a confidence interval.

If p-value $\leq \alpha$, it is often said that the null hypothesis is rejected.

Note that statistical significance does not imply practical importance.

Suppose we have a new VCR study where $n = 10000, \bar{x} = 360.2$

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{8}{\sqrt{10000}} = 0.08 \\ z &= \frac{360.2 - 360}{0.08} = 2.5 \\ p\text{−value} &= 2P[Z \geq 2.5] = 0.0124 \end{aligned}$$

So this result is statistically significant at the $\alpha = 0.05$ level. A 95% CI for μ is

$$360.2 \pm 1.96 \times 0.08 = 360.2 \pm 0.16 = (360.04, 360.36)$$

Is a difference of this size really a problem. The cost of getting closer to the desired mean of 360 minutes may be more than the cost of the extra material.

Relationship between CIs and two-sided tests:

There is a direct relationship between two-sided tests and confidence intervals. Let $\alpha = 1 - C$. Suppose that μ_0 is in the 100C% CI for μ. Then the p-value for the test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0$$

will be greater than α

$$\begin{aligned} \mu_0 \text{ in interval} &\iff \text{don't reject } H_0 \\ \mu_0 \text{ not in interval} &\iff \text{reject } H_0 \end{aligned}$$

Fixed significance level tests

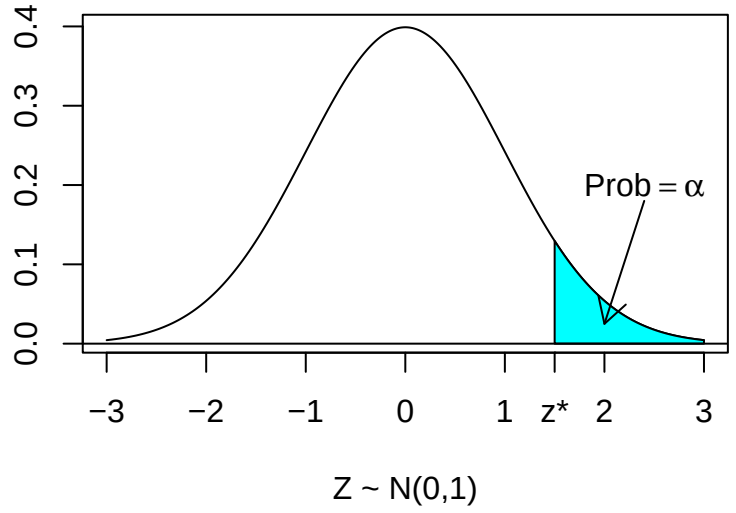
Sometimes you only want to declare statistical significance of not.

$$\text{Is } p\text{-value} \leq \alpha \text{ or is } p\text{-value} > \alpha$$

Can do this precisely without determining the p-value precisely

For a level α one-sided test

$$H_A : \mu > \mu_0 \qquad \text{Reject if } z_{obs} \geq z^*$$
$$H_A : \mu < \mu_0 \qquad \text{Reject if } z_{obs} \leq -z^*$$

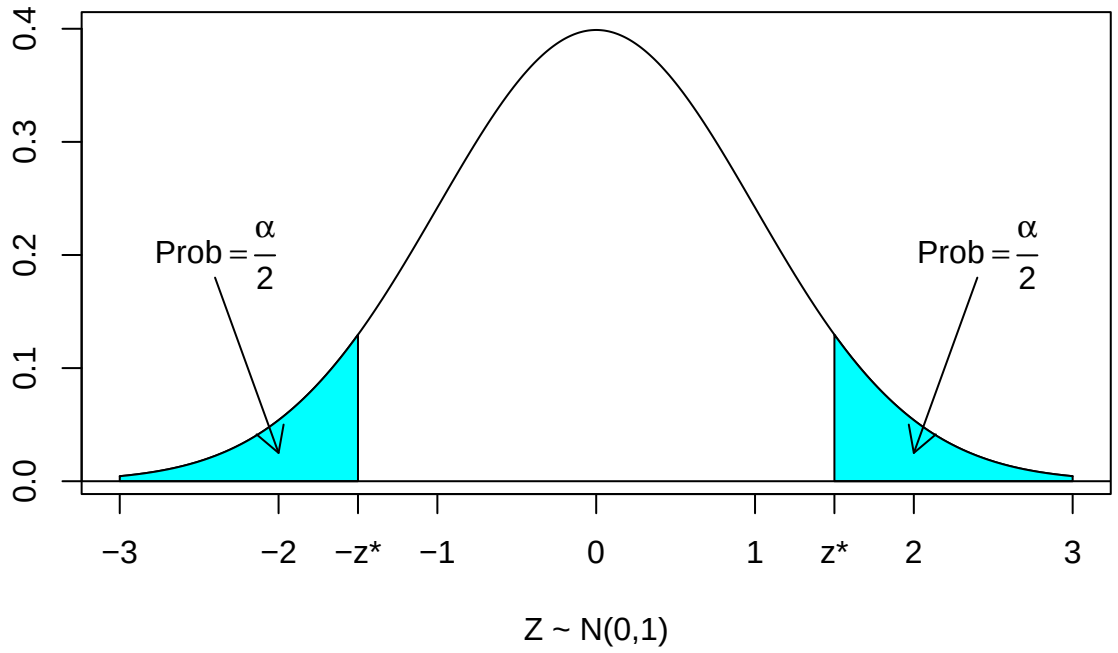


where $P[Z \geq z^*] = \alpha$

For a level α two-sided test

$$H_A : \mu \neq \mu_0 \qquad \text{Reject if } |z_{obs}| \geq z^*$$

where $P[Z \geq z^*] + P[Z \leq -z^*] = \alpha$ or $P[Z \geq z^*] = \frac{\alpha}{2}$



For these fixed level hypothesis tests, the probability that you reject H_0 when the null hypothesis test is actually true is α.

You can use Table D to get the critical values z^*.

For a one-sided test, use column α.

For a two-sided test, use column $\frac{\alpha}{2}$.

For both situations, use row z^*.

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

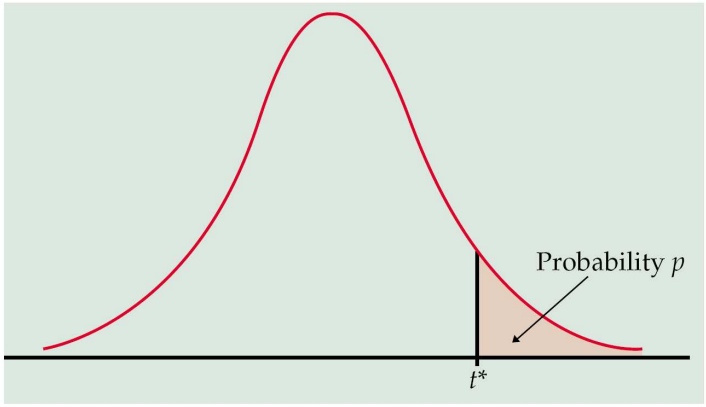


TABLE D t distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Example: Cheese making

$$H_0 : \mu = -0.545$$

$$H_A : \mu > -0.545$$

$$n = 4, \bar{x} = -0.537, \sigma_x = 0.008$$

For an $\alpha = 0.05$ test, $z^* = 1.645$.

$$z = \frac{-0.537 - (-0.545)}{\frac{0.008}{\sqrt{4}}} = \frac{0.008}{0.004} = 2$$

Since $z_{obs} = 2 > 1.645$, reject H_0 and conclude that the milk is consistent with being watered down. For comparison purposes, the

$$p\text{−value} = P[Z \geq 2] = 0.0228$$

Cautions:

The cautions mentioned for confidence intervals also apply here. In particular, we are still assuming that σ is known for the z-test, which is usually a dubious assumption. We will discuss an extension (in Section 7.1) where we can use s instead of σ in our significance test.