STATISTICAL INFERENCE

Statistical inference involves:

(a) Estimation
(b) Hypothesis Testing

Both involve using sample statistics, say, $\bar{X}$, to make inferences about the population parameter (μ).

ESTIMATION

There are two types of estimates of population parameters:

- → point estimate
- → interval estimate

A point estimate is a single number used as an estimator of a population parameter.

The problem with using a single point (or value) is that it might be right or wrong. In fact, with a continuous random variable, the probability that X is equal to a particular value is zero. [P(X=#) = 0.]

We will use an interval estimator. We say that the population parameter lies between two values.

Problem – how wide should the interval be? That depends upon how much confidence you want in the estimate.

For instance, say you wished to give a confidence interval for the mean income of a college graduate:

You might have:	that the mean income of a college grad is between
100% confidence	\$0 and \$∞
95% confidence	\$35,000 and \$41,000
90% confidence	\$36,000 and \$40,000
80% confidence	\$37,500 and \$38,500
0% confidence	\$38,000 (a point estimate)

The wider the interval, the greater the confidence you will have in it as containing the population parameter.

Confidence Interval Estimators

To estimate μ, we use:

$\bar{X} \pm Z_\alpha \sigma_{\bar{X}}$ → $\bar{X} \pm Z_\alpha \frac{\sigma}{\sqrt{n}}$ (1-α) confidence

where we get Z_α from the Z table.
[when n≥30, we use s as an estimator of σ]

To be more precise, the α error should be split in half since we are constructing a two-sided confidence interval. However, for the sake of simplicity, we will use Z_α rather than $Z_{a/2}$.

We will not have to worry about the finite population correction factor since, in most situations, n (sample size) is a very small fraction of N (population size).

EXAMPLE: Take-home salary of a NYC supermarket clerk

n = 100
$\overline{X}$ = $14,000
σ = $1,000

at 95% confidence:

$$14{,}000 \pm 1.96\ \frac{1000}{\sqrt{100}}$$

14,000 ± 196

$13,804 ←-----------------------------→ $14,196

Interpretation:
We are 95% certain that the interval from $13,804 to $14,196 contains the true population parameter, μ. Or, in 95 samples out of 100, the population mean would lie in intervals constructed by the same procedure (same n and same α).

Remember – the population parameter (μ) is fixed, it is not a random variable. Thus, it is incorrect to say that there is a 95% chance that the population mean will "fall" in this interval.

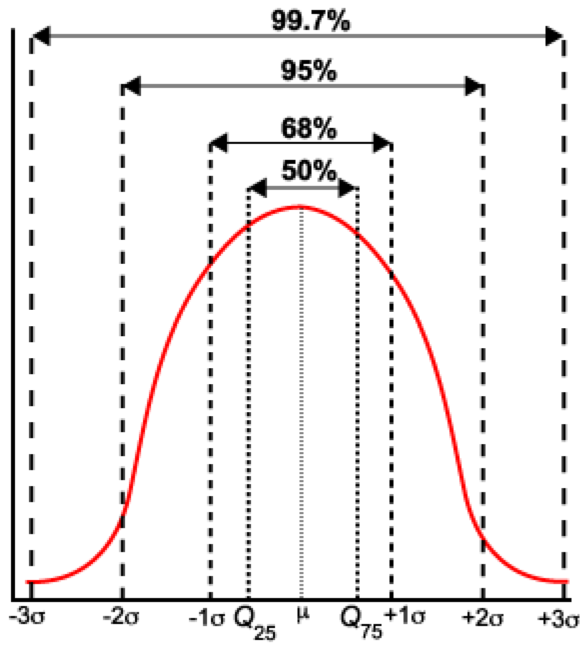
How many statisticians does it take to change a light bulb?
[one plus or minus three]

Confidence Intervals (CI)

Statistics means never having to say you're certain.

In a normal distribution:

- 68% of samples fall between ±1 SD
- 95% of samples fall between ±2 SD (actually + 1.96 SD)
- 99.7% of samples fall between ±3 SD



There is less than a 1 in 20 chance of any sample falling outside ±2 SD (95% CI, $P = 0.05$) and less than a 1 in 100 chance of any sample falling outside ±3 SD (99% CI, $P = 0.01$).

Source: http://www-micro.msb.le.ac.uk/1010/1011-18.html [This link broken - cannot find site as of Aug 7, 2012.]

EXAMPLE:

A researcher wishes to determine the average take-home pay of a part-time college student. He takes a sample of 100 students and finds:
$\overline{X}$ = \$15,000
S = \$20,000

What is the true average take home pay of a part-time college student?

(a) use a 95% confidence interval estimator.
(b) use a 99% confidence interval estimator.

In this problem we use s as an unbiased estimator of σ: E(s) = σ

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

95% Confidence Interval Estimator:

$$\overline{X} \pm Z_\alpha \, s_{\overline{X}} \quad \rightarrow \quad \overline{X} \pm Z_\alpha \frac{s}{\sqrt{n}}$$

$$15{,}000 \pm 1.96 \, \frac{20{,}000}{\sqrt{100}}$$

15,000 ± 3,920

\$11,080 ↔ \$18,920

(b) 99% confidence interval estimator:

$$15{,}000 \ \pm \ 2.575 \ \frac{20{,}000}{\sqrt{100}}$$

$$15{,}000 \ \pm \ 5{,}150$$

$$\$9{,}850 \ \leftrightarrow \ \$20{,}150$$

Note: for a given (1-α) confidence interval, there are two ways of reducing the size of the confidence interval:

1. Use a larger n. The researcher can control n.
2. Use a smaller s? Of course, this depends on the variability of the population. However, a more efficient sampling procedure (e.g., stratification) may help.

<u>EXAMPLE</u>: Average life of a GE refrigerator.

n = 100 $\quad\quad\quad\quad$ $\overline{X} \pm Z_{\alpha} \frac{s}{\sqrt{n}}$

$\overline{X}$ = 18 years

s = 4 years $\quad\quad\quad\quad$ [n≥30]

(a) <u>100% Confidence</u> $\quad\quad$ $-\infty \leftrightarrow +\infty$

$[\alpha = 0, Z_{\alpha} = \infty]$

(b) <u>99% Confidence</u>

$\alpha = .01, Z_{\alpha} = 2.575$

$18 \pm 2.575 \frac{4}{\sqrt{100}}$

18 ± 1.03 $\quad\quad$ 16.97 years ↔ 19.03 years

(c) <u>95% Confidence</u>

$\alpha = .05, Z_{\alpha} = 1.96$

$18 \pm 1.96 \frac{4}{\sqrt{100}}$

18 ± 0.78 $\quad\quad$ 17.22 years ↔ 18.78 years

(d) <u>90% Confidence</u>

$\alpha = .10, Z_{\alpha} = 1.645$

$18 \pm 1.645 \frac{4}{\sqrt{100}}$

18 ± 0.66 $\quad\quad$ 17.34 years ↔ 18.66 years

(e) 68% Confidence

$\alpha = .32, Z_{\alpha} = 1.0$

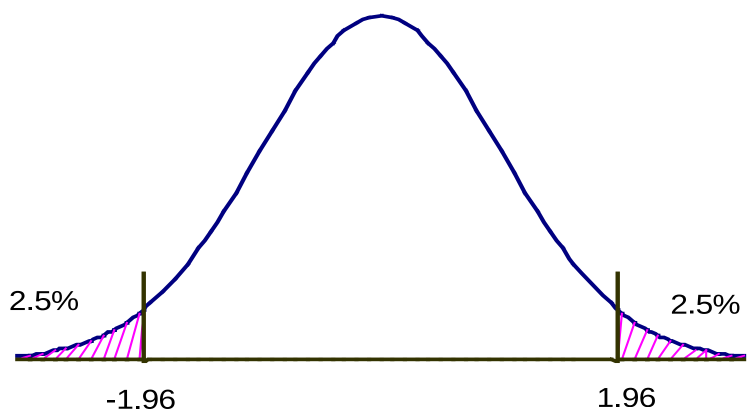
$18 \pm 1.0 \frac{4}{\sqrt{100}}$

18 ± 0.4 $\quad\quad$ 17.60 years ↔ 18.40 years

EXAMPLE

A company is interested in determining the average life of its watches (μ). An employee randomly samples 121 watches and finds that
$\overline{X}$ = 14.50 years
s = 2.00 years

We shall construct a 95% confidence interval about the mean.



The central limit theorem assures us that the sample means follow a normal distribution around the population mean when n is large. A sample size of 121 is sufficiently large. The Z value that is associated with 95% probability is +1.96 to -1.96. Thus,

$$14.50 \pm 1.96 \frac{2}{\sqrt{121}}$$

$$14.50 \pm .36$$

14.14 years ←———→ 14.86 years

The .36 is the sampling error (this is also sometimes referred to as the margin of error).

The $\frac{2}{\sqrt{121}}$ is the standard error of the mean.

By the way, given the above confidence interval, would you suggest that the company run ads stating that their watches have an average life of 15+ years? This is the kind of question that we ask when we do hypothesis testing: We test a claim about the population parameter.

<u>EXAMPLE</u>: Tire Manufacturer

n = 100
$\overline{X}$ =30,000 miles
s = 2,000 miles

Construct a 95% C.I.E of μ

$$30{,}000 \pm 1.96\ \frac{2{,}000}{\sqrt{100}}$$

30,000 ± 392

29,608 miles ↔ 30, 392 miles

<u>EXAMPLE</u>: Average Income of a College Graduate

n = 1,600
$\overline{X}$ = $50,000
s = $20,000

Construct a 99% C.I.E of μ

$$50{,}000 \pm 2.575\ \frac{20{,}000}{\sqrt{100}}$$

50,000 ± 1287.50

$48,712.50 ↔ $51,287.50

HYPOTHESIS TESTING

Suppose a company makes a claim about its product. For example, a frozen yogurt company may claim that its product has no more than 90 calories per cup. This claim is about a parameter – i.e., the population mean # calories/cup. The way the claim is tested is by taking a sample of, say, 100 cups and determining the sample mean. If the sample mean is 90 calories or less we have no evidence that the company has lied. Even if the sample mean is greater than 90 calories, it is possible the company is still telling the truth (remember sampling error). However, at some point – say, a sample average of 500 calories per cup – it will be clear that the company has no hope of claiming that they are telling the truth about their product.

A hypothesis is made about the value of some parameter, but the only facts available to estimate the true parameter are those provided by the sample. If the statistic differs from the hypothesis made about the parameter, a decision must be made as to whether or not this difference is significant. If it is, the hypothesis is rejected. If not, it cannot be rejected.

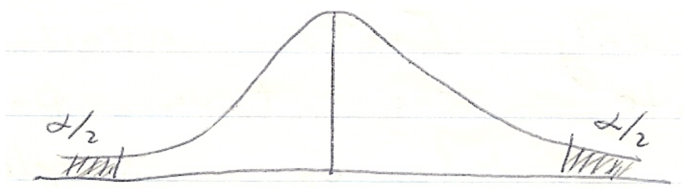
H_0 : The null hypothesis. This contains the hypothesized parameter value which will be compared with the sample value.

H_1 : The alternative hypothesis. This will be "accepted" only if H_0 is rejected.

Two types of error can occur:

		STATE OF NATURE	
		H_0 Is True	H_0 Is False
DECISION	Do Not Reject H_0	GOOD	β Error (Type II Error)
	Reject H_0	α Error (Type I Error)	GOOD

An α error occurs if we reject H_0 when it is true.



H_0: μ = #

As we try to reduce α, the possibility of making a β error increases.

There is a tradeoff.
To understand the tradeoff between the α and ß errors think of the following examples:

Legal: our legal system understands this tradeoff. If we make it extremely difficult to convict criminals because we do not want to incarcerate any innocent people we will probably have a legal system in which no one gets convicted. On the other hand, if we make it very easy to convict, then we will have a legal system in which many innocent people end up behind bars. This is why our legal system does not require a guilty verdict to be "beyond a shadow of a doubt" (i.e., complete certainty) but "beyond reasonable doubt."

Social: A woman wants to be sure that she marries the right person. She has thousands of criteria and if a suitor is missing even one trait on her list she will reject him. In statistical terms, this is a woman who is terrified of making the error of "acceptance", the error of accepting when she should reject. Unfortunately, she will probably end up rejecting a large number of suitors who would make great husbands, i.e., she will be making the error of "rejection." She has a friend who is exactly the opposite. She is terrified of not finding a husband, and therefore has virtually no criteria. She is likely to make the error of "acceptance" and very unlikely to make the error of rejection. Statisticians solve this problem by trying to limit the alpha error to some small value (say, 5%) but NOT zero.

Business: a company purchases chips for its computers. It purchases them in batches of 1,000. The company is willing to live with a few defects per 1,000 chips. How many defects? If it randomly samples 100 chips from each batch and rejects the entire shipment if there are ANY defects, they may end up rejecting too many shipments. Of course, if they are too liberal in what they accept and assume everything is "sampling error," they are very likely to make the error of acceptance.

All of this is very similar (in fact exactly the same) as the problem we had earlier with confidence intervals. Ideally, we would love a very narrow interval, with a lot of confidence. But, practically, we can never have both: there is a tradeoff.

We can actually test a hypothesis using the confidence interval estimators we already learned.

EXAMPLE:
The Duracell Battery Company claims that its new Bunnywabbit batteries have a life, on the average, of 1,000 hours.

Suppose, you take a sample of 100 batteries and test them. You find:
$\overline{X}$ = 985 hours
s = 30 hours

Construct a 95% CIE and decide whether the company's claim should be rejected or not.

$985 \pm 1.96 \frac{30}{\sqrt{100}}$ hours

985 ± 5.88 hours

979.1 hours ↔ 990.9 hours

You are 95% sure that the true μ is somewhere between ["is covered by the interval"] 979.1 and 990.9 hours.

Notice that if we were to standardize the sample average of 985 …

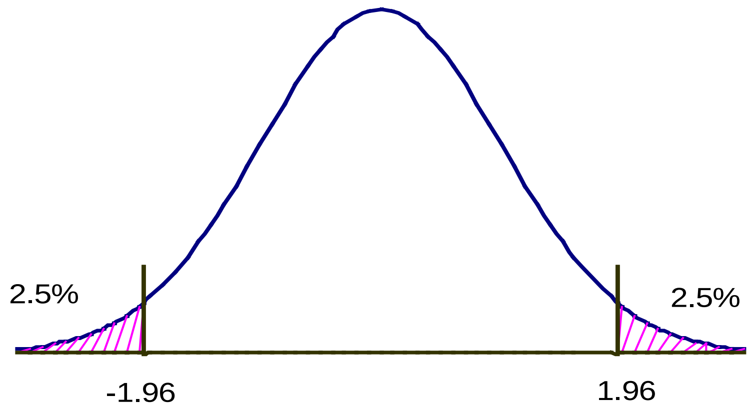
$$Z = \frac{985 - 1000}{\frac{30}{\sqrt{100}}} = -15/3 = -5$$

… we would find that it is 5 standard deviations away from the mean. How likely is this?

This is equivalent to testing at the .05 level.

H_0: μ = 1,000 hours
H_1: μ ≠ 1,000 hours

REJECT THE CLAIM [H_0]



[EXPLAIN – region of "acceptance" and region of rejection]

Steps in Hypothesis Testing

You test a hypothesis—an assertion or claim about a parameter by using sample evidence. The sample evidence is converted into a Z-score – in other words, it is standardized – using the hypothesized μ value.

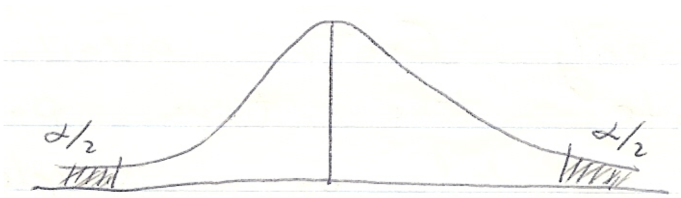
$$Z = \frac{\overline{X} - \mu_H}{\sigma / \sqrt{n}}$$

If n is large, s can be used in lieu of σ.

H_0 is a hypothesis about the value of some parameter.

1. Formulate H_0 and H_1. H_0 is the null hypothesis and H_1 is the alternative hypothesis.
2. Specify the level of significance (α) to be used. This level of significance tells you the probability of rejecting H_0 when it is, in fact, true. (Normally, significance level of 0.05 or 0.01 are used)
3. Select the test statistic: e.g., Z, t, χ^2, F, etc.
4. Establish the critical value or values of the test statistic needed to reject H_0. DRAW A PICTURE!
5. Determine the actual value (computed value) of the test statistic.
6. Make a decision: **Reject H_0 or Do Not Reject H_0**.

If you recall, we used Z_α when we constructed confidence intervals. The α refers to the alpha error above. We are, say, 95% certain that the interval we created contains the population mean. This suggests that there is a 5% chance that the interval does not contain the population mean, i.e., we have made an error.

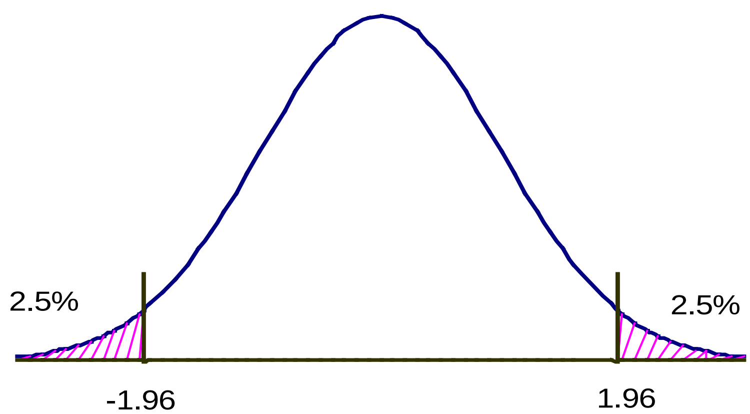


H_0: $\mu = \#$

With a "two-tail" hypothesis test, α is split into two and put in both tails. H1 contains two possibilities: $\mu > \#$ OR $\mu < \#$. This is why the region of rejection is divided into two tails. Note that the region of rejection always corresponds to H_1.

EXAMPLE:
A pharmaceutical company claims that each of its pills contains exactly 20.00 milligrams of Cumidin (a blood thinner). You sample 64 pills and find that $\overline{X}$ =20.50 mgs. and s = .80 mgs. Should the company's claim be rejected? Test at $\alpha = 0.05$.

H_0: $\mu = 20.00$ mgs
H_1: $\mu \neq 20.00$ mgs



$$Z = \frac{20.50 - 20.00}{\frac{.80}{\sqrt{64}}} = \frac{.50}{.10} = 5$$

[$\frac{.80}{\sqrt{64}}$ = .10 This is the standard error of the mean.]

The Z value of 5 is deep in the rejection region.

Therefore, reject H_0 at $p < .05$

If we took the above data and constructed a 95% confidence interval:

95%, Confidence Interval = 20.50 ± 1.96(.10)

20.304 mg ←———→ 20.696 mg

Note that 20.00 mg is not in this interval.

The hypothesis test in this example is called a Two-tail Test, because the region of rejection is split (equally) into the two tails of the distribution.

When you do a two-tail test at an alpha of .05 (.05 significance level) you will come to exactly the same conclusions as when you construct a two-sided 95% confidence interval. Indeed, you will be using the same Z-values, ± 1.96.

So, if we can test hypotheses using confidence interval estimate, why don't we just do that? There is an additional piece of information that we can get out of the hypothesis test, and that you get from the MS Excel output: the p value.