Introduction to Inference

IPS Chapter 6

- 6.1: Estimating with Confidence
- 6.2: Tests of Significance
- 6.3: Use and Abuse of Tests
- 6.4: Power and Inference as a Decision

© 2012 W.H. Freeman and Company

Objectives

6.1 Estimating with confidence

- Statistical confidence
- Confidence intervals
- Confidence interval for a population mean
- How confidence intervals behave
- Choosing the sample size

Overview of Inference

- Methods for drawing conclusions about a population from sample data are called statistical inference

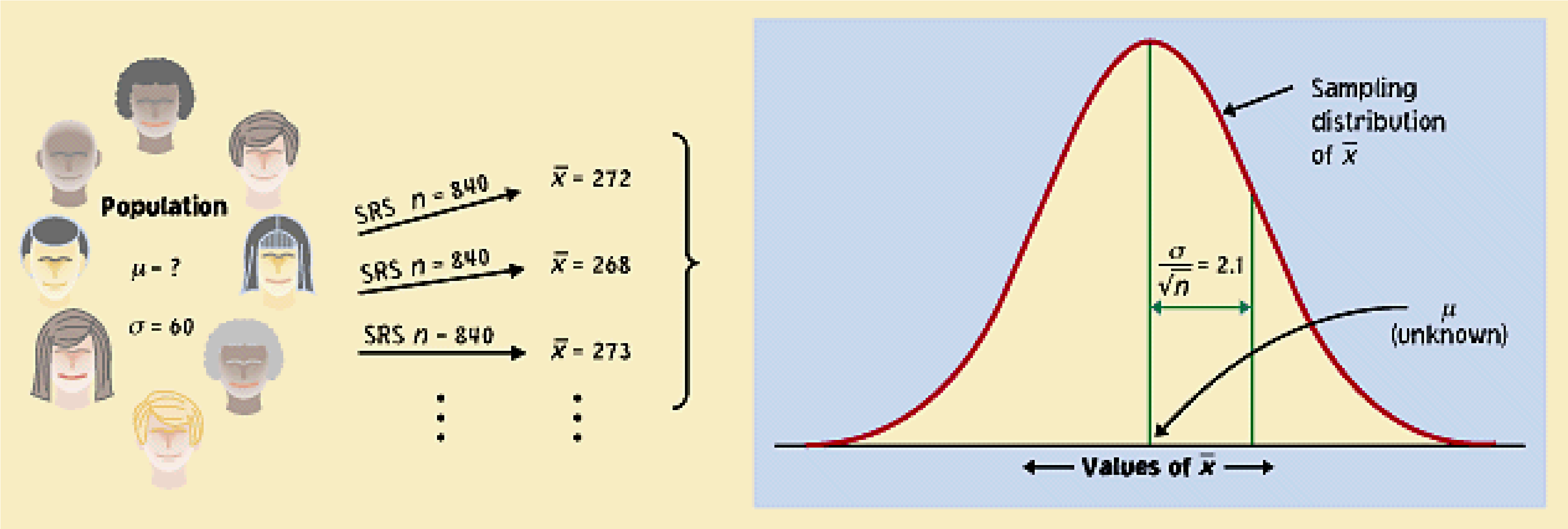
- Methods
  - Confidence Intervals - estimating a value of a population parameter
  - Tests of significance - assess evidence for a claim about a population

- Inference is appropriate when data are produced by either
  - a random sample or
  - a randomized experiment

Statistical confidence

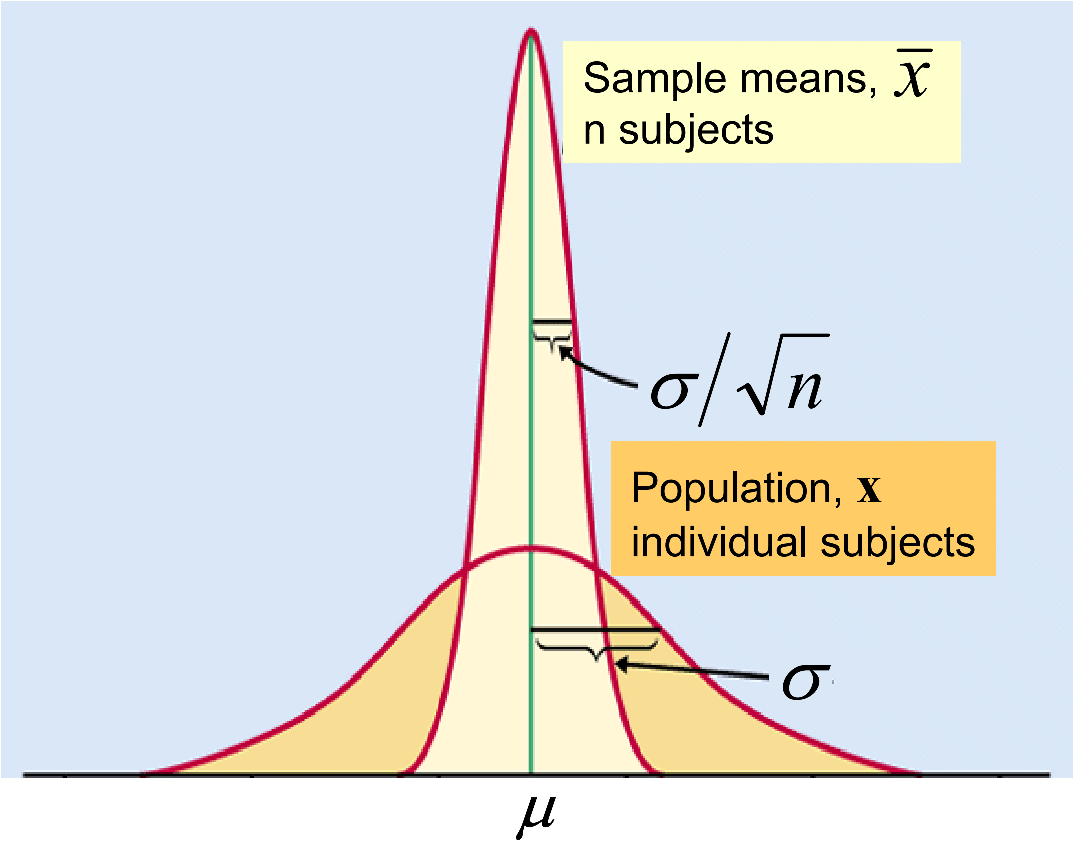
Although the sample mean, $\overline{x}$, is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean.

In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ.

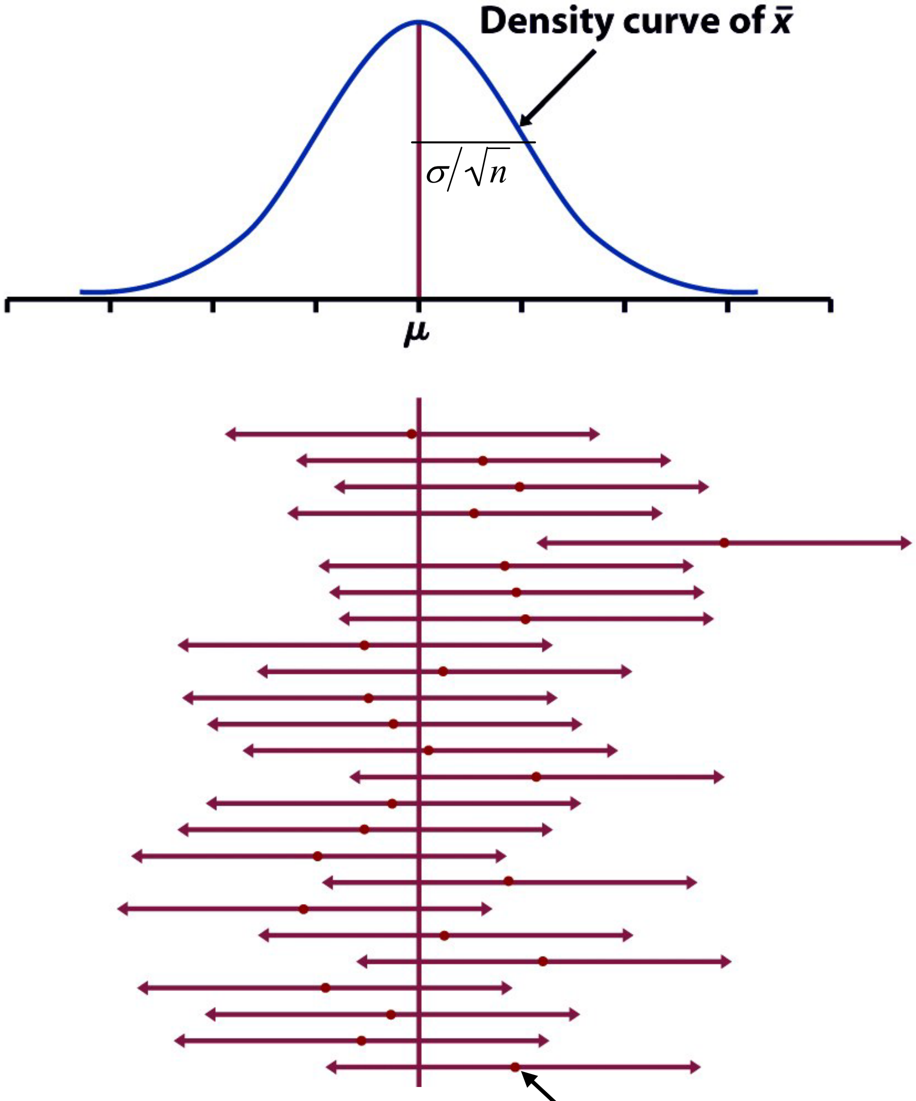


But the sample distribution is narrower than the population distribution, by a factor of $\sqrt{n}$.

Thus, the estimates $\bar{x}$ gained from our samples are always relatively close to the population parameter μ.



If the population is normally distributed $N(\mu, \sigma)$, so will the sampling distribution $N(\mu, \sigma/\sqrt{n})$, 95% of all sample means will be within roughly 2 standard deviations ($2*\sigma/\sqrt{n}$) of the population parameter μ.

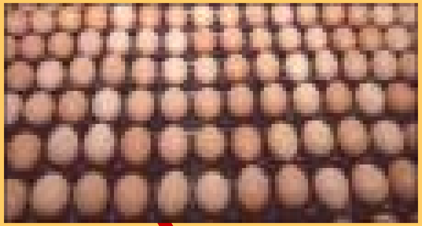


Red dot: mean value of individual sample

Distances are symmetrical which implies that **the population parameter μ must be within roughly 2 standard deviations from the sample average $\bar{x}$, in 95% of all samples.**

This reasoning is the essence of statistical inference.

The weight of single eggs of the brown variety is normally distributed *N*(65 g,5 g). Think of a carton of 12 brown eggs as an SRS of size 12.

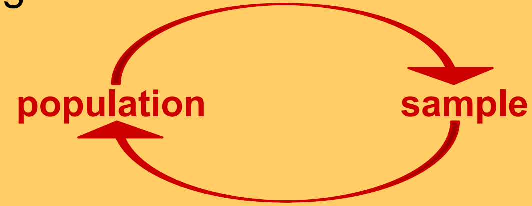


- What is the distribution of the sample means $\bar{x}$?

  Normal (mean μ, standard deviation $\sigma/\sqrt{n}$) = *N*(65 g,1.44 g).

- Find the middle 95% of the sample means distribution.

  Roughly ± 2 standard deviations from the mean, or 65g ± 2.88g.



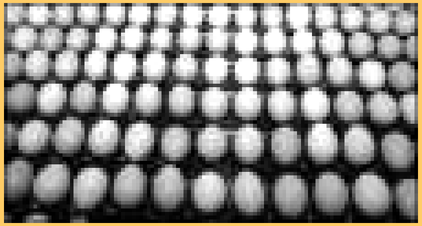
population

sample

You buy a carton of 12 white eggs instead. The box weighs 770 g. The average egg weight from that SRS is thus $\bar{x}$ = 64.2 g.

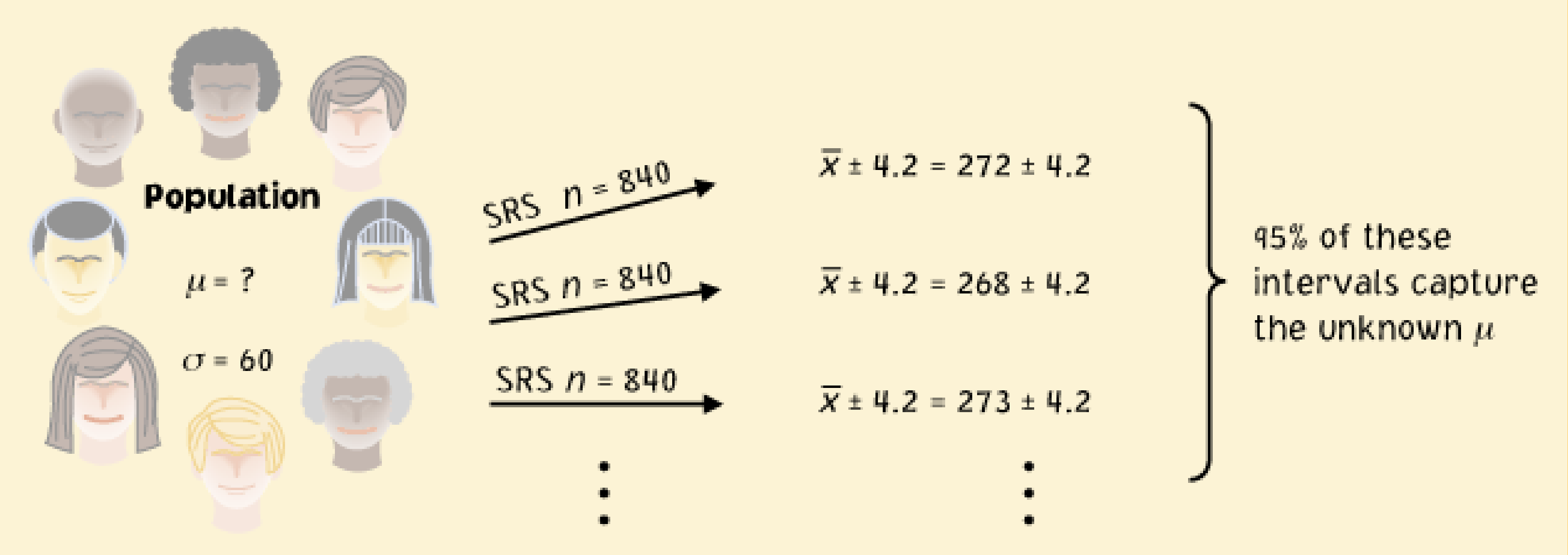
- Knowing that the standard deviation of egg weight is 5 g, what can you infer about the mean μ of the white egg population?

  We are 95% confident that the population mean μ is between 64.2 g ± 2.88 g, or roughly within ± $2\sigma/\sqrt{n}$ of $\bar{x}$.



Confidence intervals

The **confidence interval** is a range of values with an associated probability or **confidence level *C***. The probability quantifies the chance that the interval contains the true population parameter.

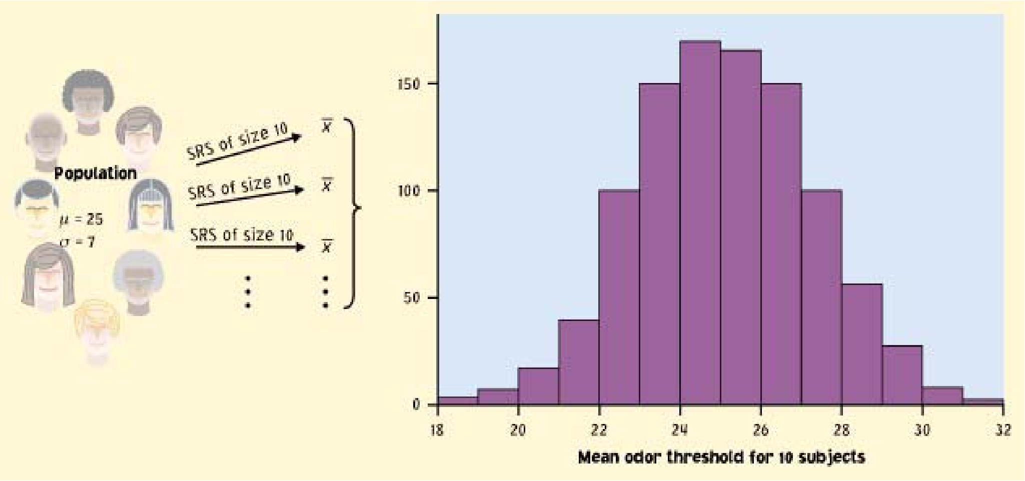

Population
μ = ?
σ = 60
SRS n = 840
SRS n = 840
SRS n = 840
$\bar{x}$ ± 4.2 = 272 ± 4.2
$\bar{x}$ ± 4.2 = 268 ± 4.2
$\bar{x}$ ± 4.2 = 273 ± 4.2
95% of these intervals capture the unknown μ

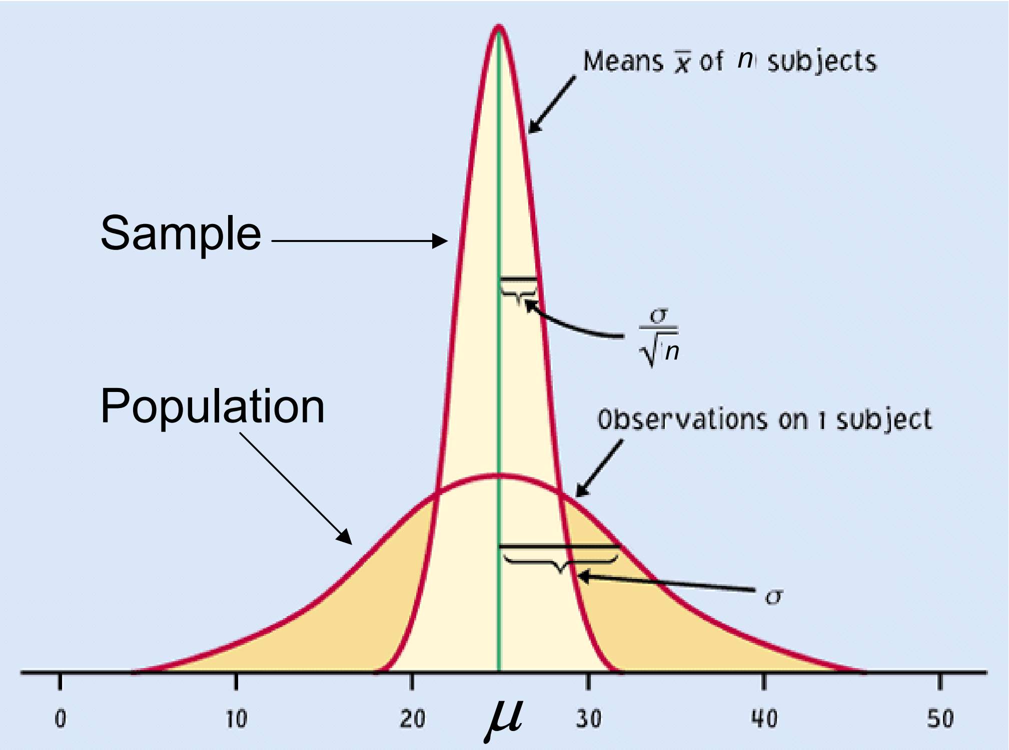
$\bar{x}$ ± 4.2 is a 95% confidence interval for the population parameter μ.

This equation says that in 95% of the cases, the actual value of μ will be within 4.2 units of the value of $\bar{x}$.

Implications

We don’t need to take a lot of random samples to “rebuild” the sampling distribution and find μ at its center.



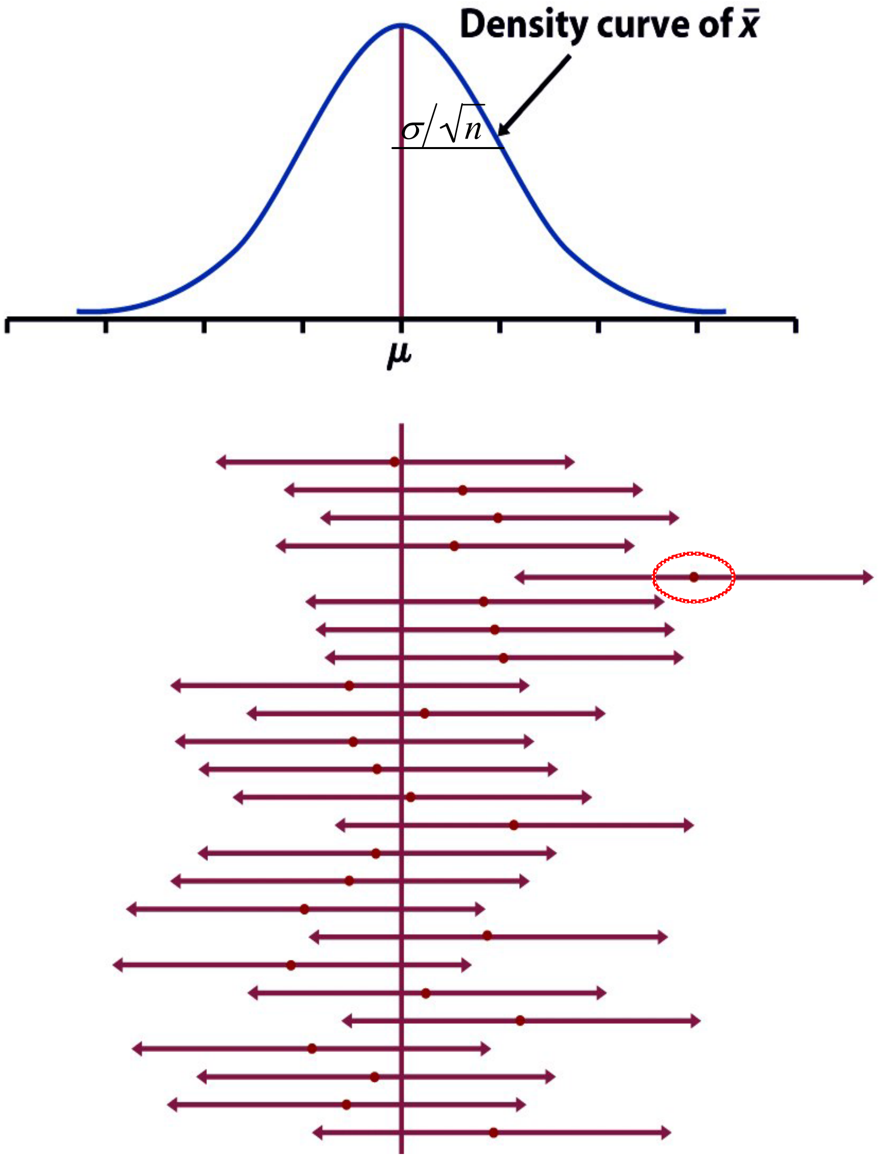

Means $\bar{x}$ of n subjects
Sample
$\frac{\sigma}{\sqrt{n}}$
Population
Observations on 1 subject
σ
0
10
20
μ
30
40
50

All we need is one SRS of size *n* and rely on the properties of the sample means distribution to infer the population mean μ.

Reworded

With 95% confidence, we can say that μ should be within roughly 2 standard deviations (2*$\sigma/\sqrt{n}$) from our sample mean $\bar{x}$.

- In 95% of all possible samples of this size n, μ will indeed fall in our confidence interval.
- In only 5% of samples would $\bar{x}$ be farther from μ.



A **confidence interval** can be expressed as:

- Mean ± *m*

  m is called the **margin of error**

  μ within $\bar{x} \pm m$

  Example: 120 ± 6

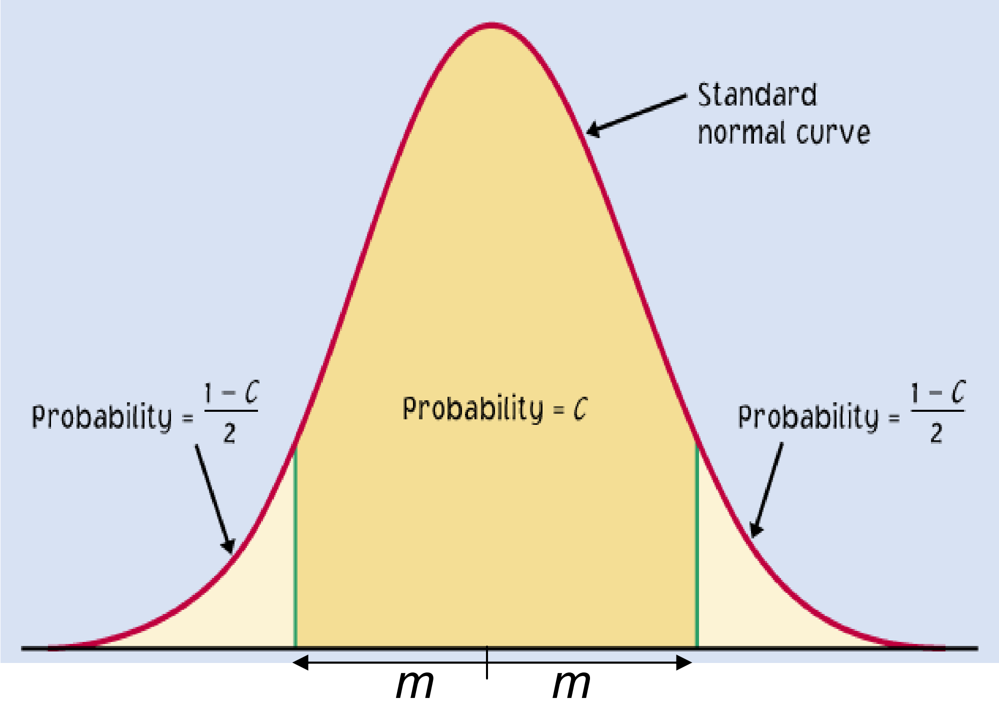
- Two endpoints of an interval

  μ within $(\bar{x} - m)$ to $(\bar{x} + m)$

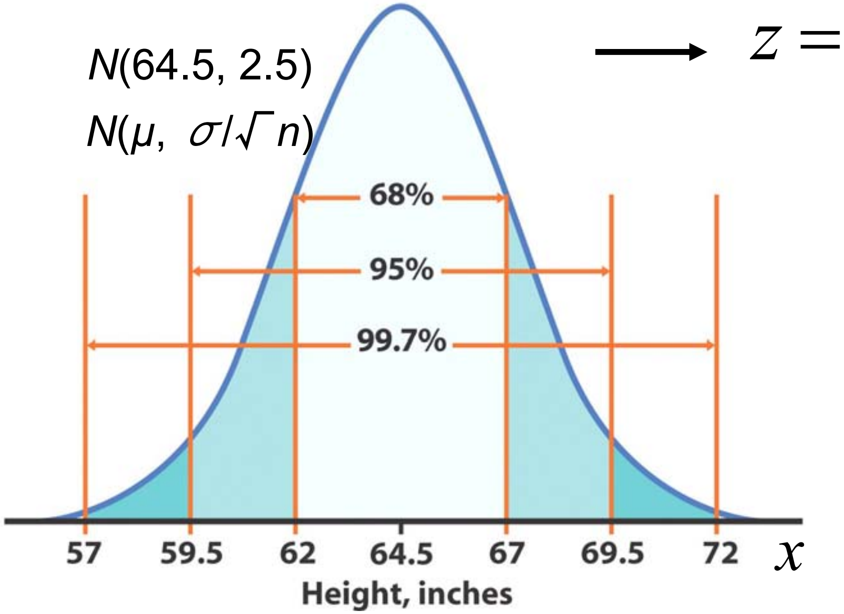
  ex. 114 to 126

A **confidence level *C*** (in %) indicates the probability that the μ falls within the interval.

It represents the area under the normal curve within ± *m* of the center of the curve.



Review: standardizing the normal curve using *z*



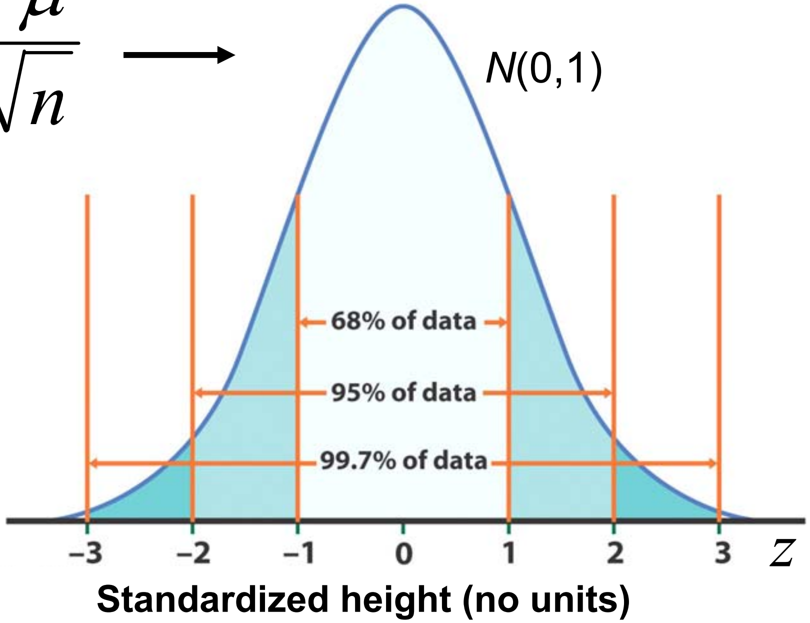
N(64.5, 2.5)

N(μ, $\sigma/\sqrt{n}$)

→

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

→



N(0,1)

Here, we work with the sampling distribution,
and $\sigma/\sqrt{n}$ is its standard deviation (spread).

Remember that σ is the standard deviation of the original population.

Varying confidence levels

Confidence intervals contain the population mean μ in *C*% of samples.

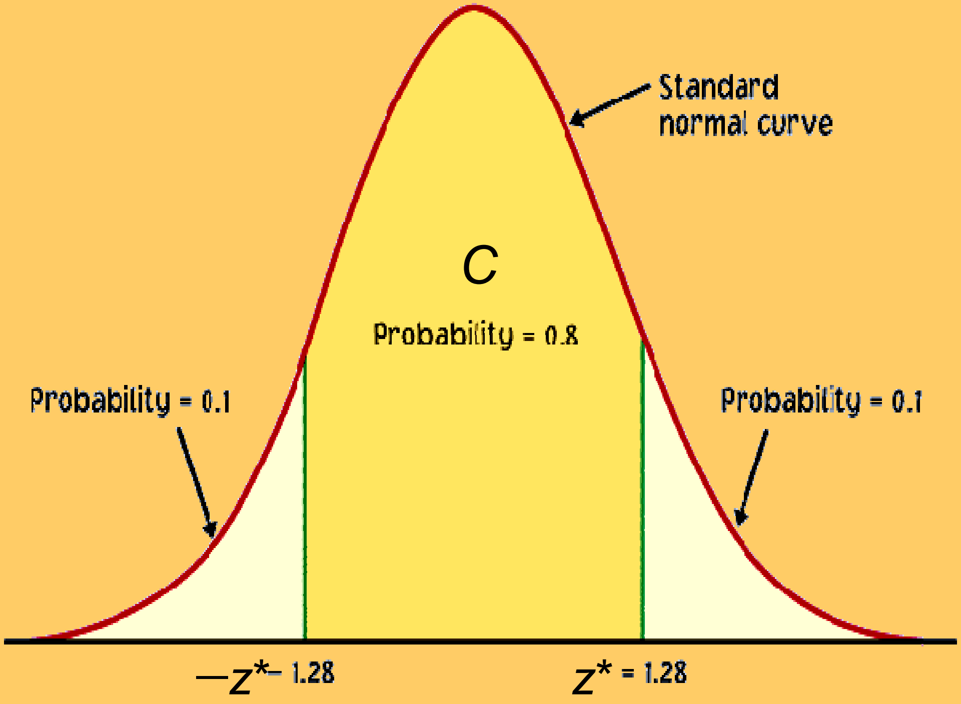
Different areas under the curve give different confidence levels *C*.

Practical use of z: z*

- *z** is related to the chosen confidence level *C*.
- *C* is the area under the standard normal curve between −*z** and *z**.

The confidence interval is thus:

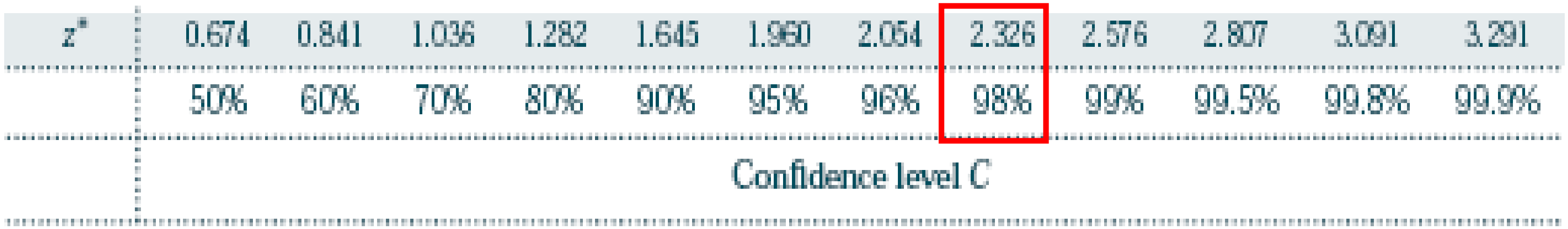
$$\bar{x} \pm z^* \sigma / \sqrt{n}$$



Example: For an 80% confidence level *C*, 80% of the normal curve's area is contained in the interval.

How do we find specific *z** values?

We can use a table of *z/t* values (Table D). For a particular confidence level, *C*, the appropriate *z** value is just above it.



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: For a 98% confidence level, *z**=2.326

We can use software. In **Excel:**

=NORMINV(probability,mean,standard_dev)

gives z for a given cumulative probability.

Since we want the middle *C* probability, the probability we require is (1 - C)/2

Example: For a 98% confidence level, =NORMINV(.01,0,1) = −2.32635 (= neg. *z**)

Link between confidence level and margin of error

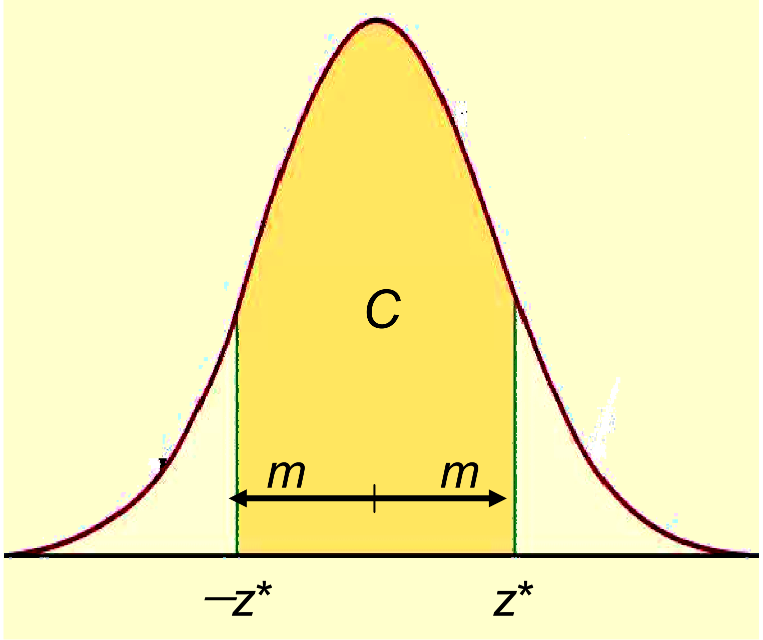
The confidence level *C* determines the value of *z** (in table C).

The margin of error also depends on *z**.

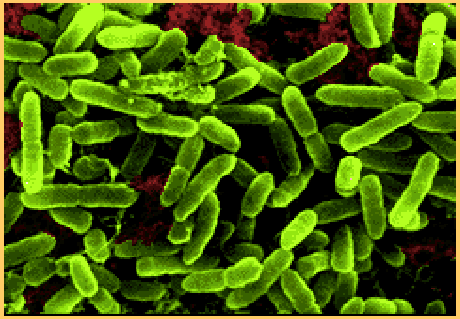
$$m = z^* \sigma / \sqrt{n}$$

Higher confidence ***C*** implies a larger margin of error ***m*** (thus less precision in our estimates).

A smaller confidence level ***C*** produces a smaller margin of error ***m*** (thus better precision in our estimates).



Different confidence intervals for the same set of measurements



Density of bacteria in solution:

Measurement equipment has standard deviation $\sigma = 1 * 10^6$ bacteria/ml fluid.

Three measurements: 24, 29, and 31 * 10^6 bacteria/ml fluid

Mean: $\bar{x} = 28 * 10^6$ bacteria/ml. Find the 96% and 70% CI.

- 96% confidence interval for the true density, $z^* = 2.054$, and write

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 2.054(1/\sqrt{3})$$

$$= 28 \pm 1.19 \times 10^6$$

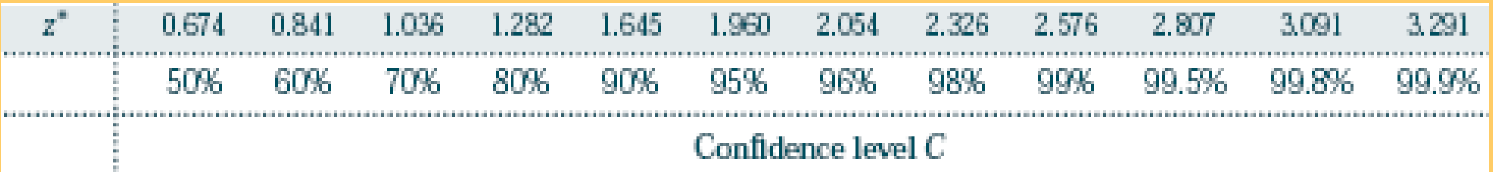
bacteria/ml

- 70% confidence interval for the true density, $z^* = 1.036$, and write

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 1.036(1/\sqrt{3})$$

$$= 28 \pm 0.60 \times 10^6$$

bacteria/ml



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											

Properties of Confidence Intervals

- User chooses the confidence level
- Margin of error follows from this choice

We want

- high confidence
- small margins of error

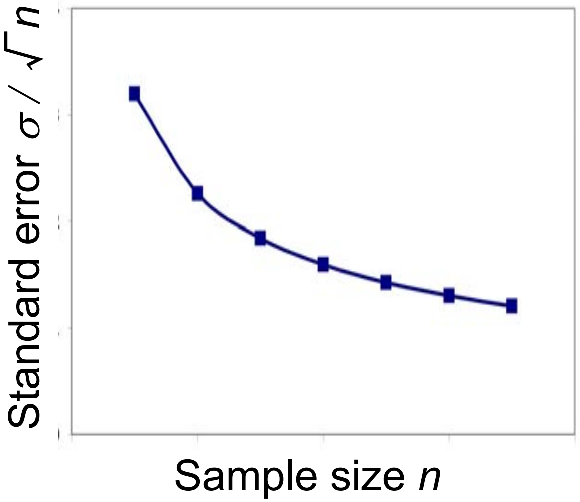
The margin of error, $z^{*}\sigma/\sqrt{n}$, gets smaller when

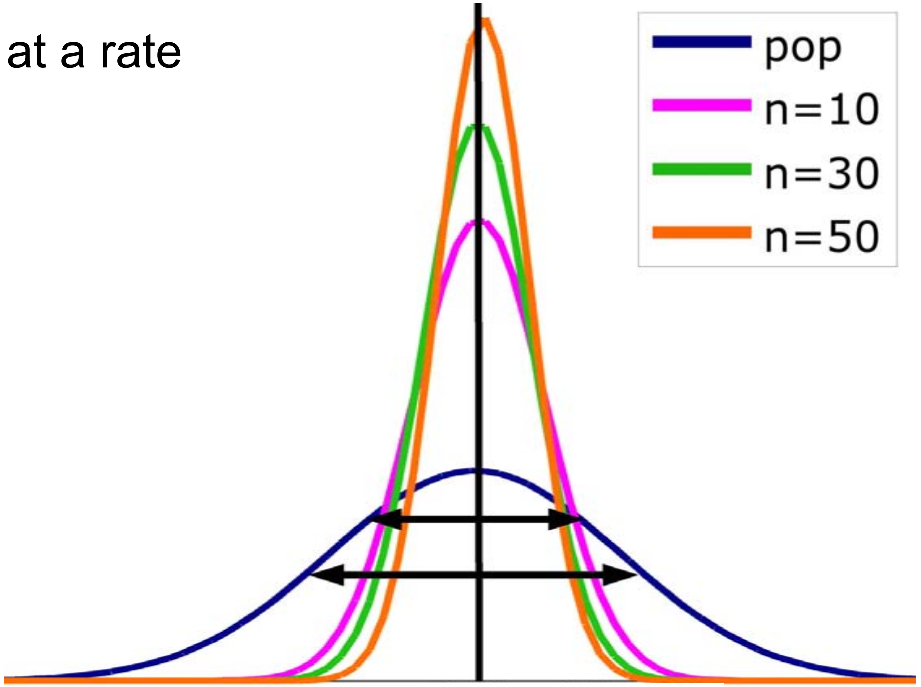
- z* (and thus the confidence level C) gets smaller
- σ is smaller
- n is larger

Impact of sample size

The spread in the sampling distribution of the mean is a function of the number of individuals per sample.

- The larger the sample size, the smaller the standard deviation (spread) of the sample mean distribution.
- But the spread only decreases at a rate equal to $\sqrt{n}$.





Sample size and experimental design

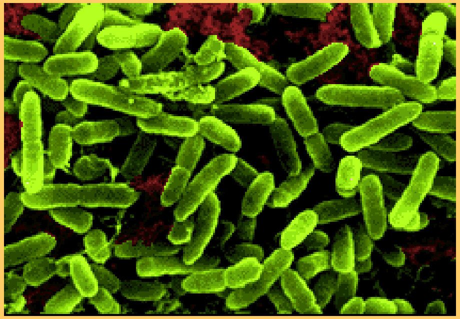
You may need a certain margin of error (e.g., drug trial, manufacturing specs). In many cases, the population variability (σ) is fixed, but we can choose the number of measurements (n).

So plan ahead what sample size to use to achieve that margin of error.

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad n = \left(\frac{z^* \sigma}{m} \right)^2$$

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples. The best approach is to use the smallest sample size that can give you useful results.

What sample size for a given margin of error?



Density of bacteria in solution:

Measurement equipment has standard deviation

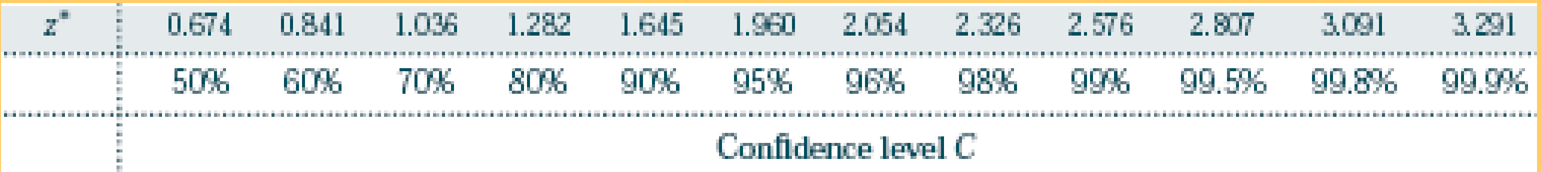
σ = 1 * 10^6 bacteria/ml fluid.

How many measurements should you make to obtain a margin of error of at most 0.5 * 10^6 bacteria/ml with a confidence level of 95%?

For a 95% confidence interval, *z** = 1.96.

$$n = \left(\frac{z^* \sigma}{m}\right)^2 \quad \Rightarrow \quad n = \left(\frac{1.96 * 1}{0.5}\right)^2 = 3.92^2 = 15.3664$$

Using only 15 measurements will not be enough to ensure that *m* is no more than 0.5 * 10^6. Therefore, we need at least 16 measurements.



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											

Cautions about using $\bar{x} \pm z^{*} * \sigma/\sqrt{n}$

- Data must be a **SRS** from the population.
- Formula is **not** correct for other sampling designs.
- Inference **cannot** rescue badly produced data.
- Confidence intervals are **not resistant** to outliers.
- If ***n*** is small **(<15)** and the population is not normal, the true confidence level will be **different** from ***C***.
- The standard deviation σ of the population must be known.

◆ **The margin of error in a confidence interval covers only random sampling errors!**

Interpretation of Confidence Intervals

- Conditions under which an inference method is valid are **never fully met in practice**. Exploratory data analysis and judgment should be used when deciding whether or not to use a statistical procedure.

- Any individual confidence interval either **will or will not** contain the true population mean. It is **wrong** to say that the probability is 95% that the true mean falls in the confidence interval.

- The **correct interpretation** of a 95% confidence interval is that we are 95% confident that the true mean falls within the interval. The confidence interval was calculated by a method that gives correct results in 95% of all possible samples.

  In other words, if many such confidence intervals were constructed, 95% of these intervals would contain the true mean.

Introduction to Inference

6.2 Tests of Significance

© 2012 W.H. Freeman and Company

Objectives

6.2 Tests of significance

- The reasoning of significance tests
- Stating hypotheses
- The *P*-value
- Statistical significance
- Tests for a population mean
- Confidence intervals to test hypotheses

Reasoning of Significance Tests

We have seen that the properties of the sampling distribution of $\bar{x}$ help us estimate a range of likely values for population mean μ.

We can also rely on the properties of the sample distribution to test hypotheses.

Example: You are in charge of quality control in your food company. You sample randomly four packs of cherry tomatoes, each labeled 1/2 lb. (227 g).

The average weight from your four boxes is 222 g. Obviously, we cannot expect boxes filled with whole tomatoes to all weigh exactly half a pound. Thus,

- Is the somewhat smaller weight simply due to chance variation?
- Is it evidence that the calibrating machine that sorts cherry tomatoes into packs needs revision?

Stating hypotheses

A **test of statistical significance** tests a specific hypothesis using sample data to decide on the validity of the hypothesis.

In statistics, a **hypothesis** is an assumption or a theory about the characteristics of one or more variables in one or more populations.

What you want to know: Does the calibrating machine that sorts cherry tomatoes into packs need revision?

The same question reframed statistically: Is the population mean μ for the distribution of weights of cherry tomato packages equal to 227 g (i.e., half a pound)?

The **null hypothesis** is a very specific statement about a parameter of the population(s). It is labeled H_0.

The **alternative hypothesis** is a more general statement about a parameter of the population(s) that is exclusive of the null hypothesis. It is labeled H_a.

Weight of cherry tomato packs:

$H_0 : \mu = 227$ g (μ is the average weight of the population of packs)

$H_a : \mu \neq 227$ g (μ is either larger or smaller)

One-sided and two-sided tests

- A **two-tail** or **two-sided test** of the population mean has these null and alternative hypotheses:

$$H_0:\ \mu = \text{[a specific number]} \quad H_a:\ \mu \neq \text{[a specific number]}$$

- A **one-tail** or **one-sided test** of a population mean has these null and alternative hypotheses:

$$H_0:\ \mu = \text{[a specific number]} \quad H_a:\ \mu < \text{[a specific number]} \qquad \text{OR}$$

$$H_0:\ \mu = \text{[a specific number]} \quad H_a:\ \mu > \text{[a specific number]}$$

The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug it is copying. Higher or lower absorption would both be problematic, thus we test:

$$H_0: \mu_{\text{generic}} = \mu_{\text{brand}} \qquad H_a: \mu_{\text{generic}} \neq \mu_{\text{brand}} \qquad \text{two-sided}$$

How to choose?

What determines the choice of a one-sided versus a two-sided test is what we know about the problem before we perform a test of statistical significance.

A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg.

Here, the health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise in order to better addict consumers to their products and maintain revenues.

Thus, this is a one-sided test: $H_0 : \mu = 1.4$ mg $H_a : \mu > 1.4$ mg

It is important to make that choice before performing the test or else you could make a choice of “convenience” or fall into circular logic.

The P-value

The packaging process has a known standard deviation $\sigma = 5$ g.

$H_0 : \mu = 227$ g versus $H_a : \mu \neq 227$ g

The average weight from your four random boxes is 222 g.

What is the probability of drawing a random sample such as yours if H_0 is true?

Tests of statistical significance quantify the chance of obtaining a particular random sample result if the null hypothesis were true. This quantity is the **P-value.**

This is a way of assessing the "believability" of the null hypothesis, given the evidence provided by a random sample.

Interpreting a P-value

Could random variation alone account for the difference between the null hypothesis and observations from a random sample?

- A small P-value implies that random variation due to the sampling process alone is not likely to account for the observed difference.

- With a small p-value we **reject H_0.** The true property of the population is **significantly** different from what was stated in H_0.

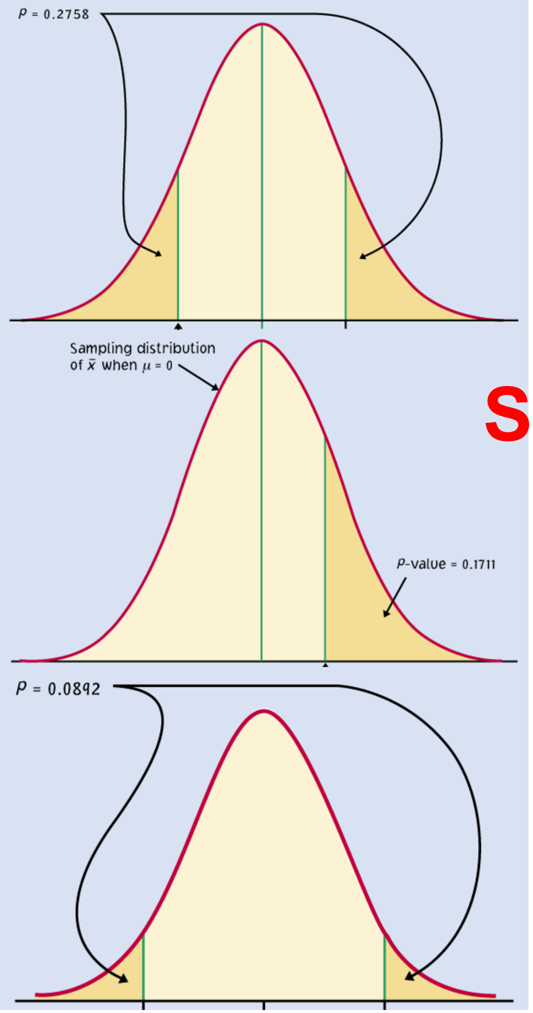
Thus, small P-values are strong evidence AGAINST H_0.

But how small is small…?

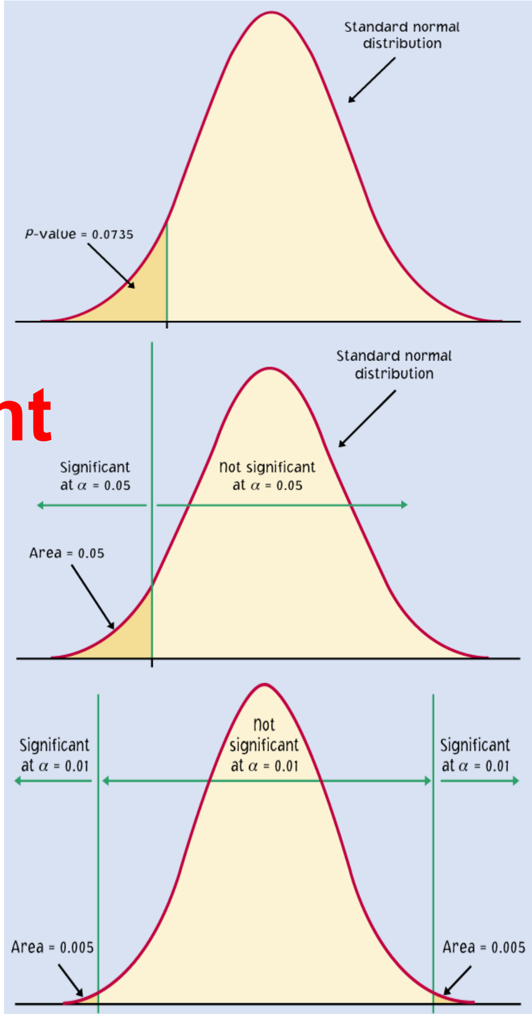
P = 0.2758

P = 0.1711

P = 0.0892



Significant P-value ???


Standard normal distribution
P-value = 0.0735
Standard normal distribution
Significant at α = 0.05
Not significant at α = 0.05
Area = 0.05
Significant at α = 0.01
Not significant at α = 0.01
Significant at α = 0.01
Area = 0.005
Area = 0.005

P = 0.0735

P = 0.05

P = 0.01

When the shaded area becomes very small, the probability of drawing such a sample at random gets very slim. Oftentimes, a P-value of 0.05 or less is considered **significant:** The phenomenon observed is unlikely to be entirely due to chance event from the random sampling.

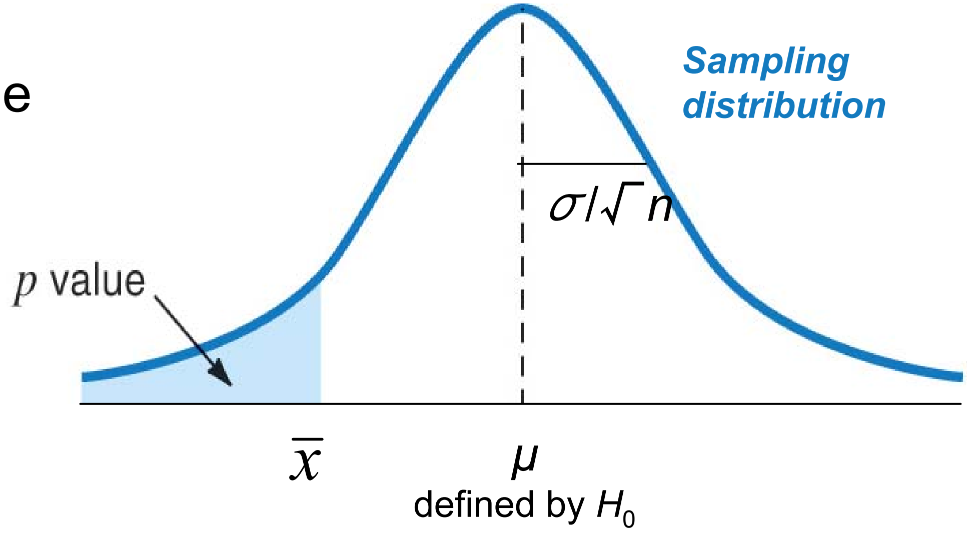
Tests for a population mean

To test the hypothesis $H_0 : \mu = \mu_0$ based on an SRS of size n from a Normal population with unknown mean μ and known standard deviation σ, we rely on the properties of the sampling distribution $N(\mu, \sigma/\sqrt{n})$.

The P-value is the area under the sampling distribution for values at least as extreme, in the direction of H_a, as that of our random sample.

Again, we first calculate a z-value and then use Table A.

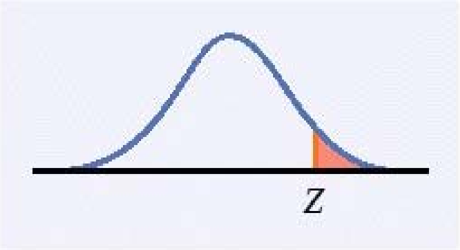
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$



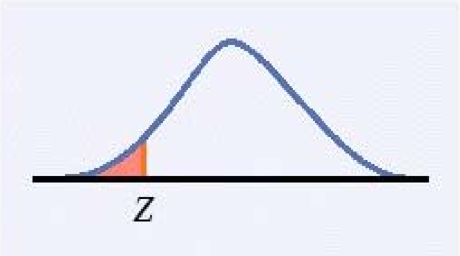
P-value in one-sided and two-sided tests

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

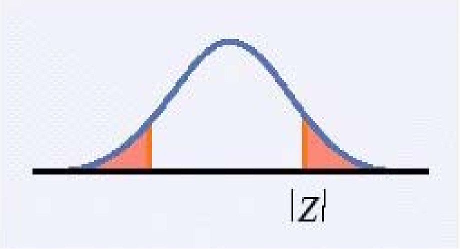


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.

Does the packaging machine need revision?

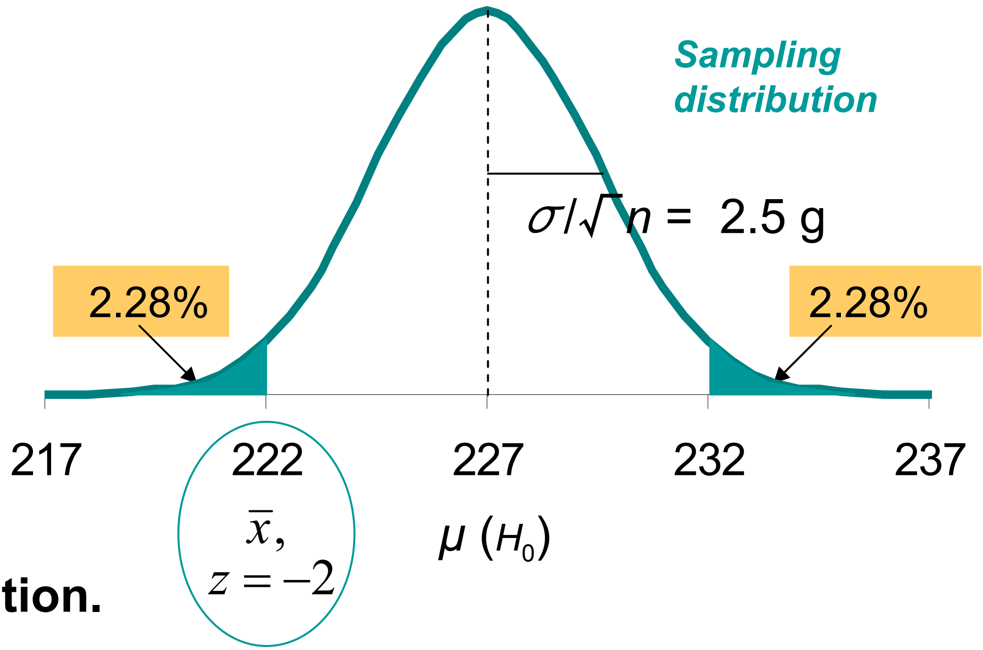
- $H_0 : \mu = 227$ g versus $H_a : \mu \neq 227$ g
- What is the probability of drawing a random sample such as yours if H_0 is true?

$$\bar{x} = 222\text{g} \qquad \sigma = 5\text{g} \qquad n = 4$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{222 - 227}{5 / \sqrt{4}} = -2$$

From table A, the area under the standard normal curve to the left of z is 0.0228.

Thus, P-value = 2*0.0228 = 4.56%.



The probability of getting a random sample average so different from μ is so low that we reject H_0.

→**The machine does need recalibration.**

Steps for Tests of Significance

1. State the *null hypotheses* H_o and the *alternative hypothesis* H_a.

2. Calculate value of the *test statistic*.

3. Determine the *P-value* for the observed data.

4. State a conclusion.

The significance level: α

The significance level, α, is the largest P-value tolerated for rejecting a true null hypothesis (how much evidence against H_0 we require). This value is decided arbitrarily <u>before</u> conducting the test.

- If the P-value is equal to or less than α (**P $\leq$ α**), then we **reject** ***H_0***.
- If the P-value is greater than α (**P > α**), then we **fail to reject *H***

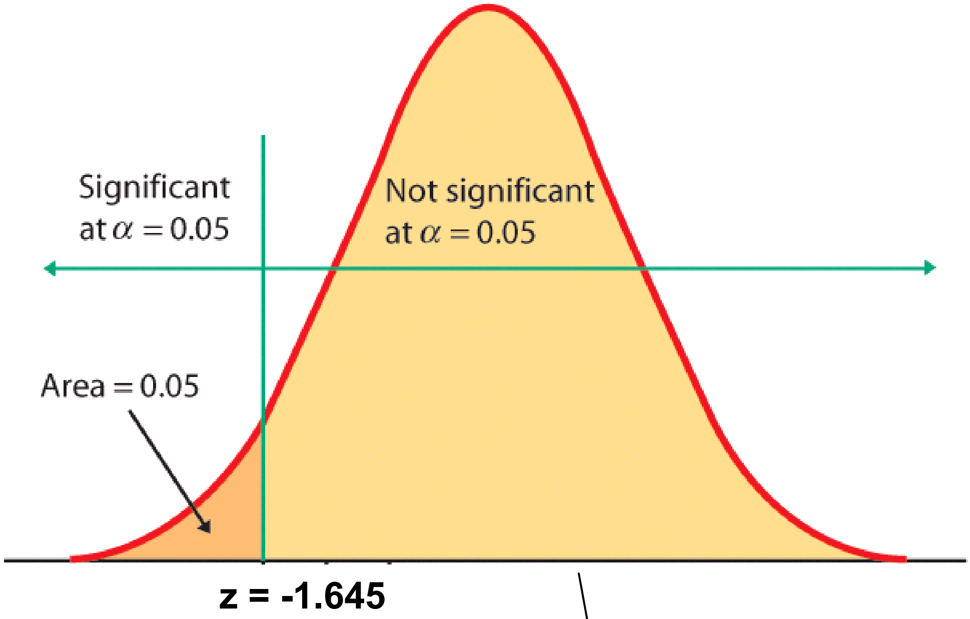
Does the packaging machine need revision?

Two-sided test. The P-value is 4.56%.

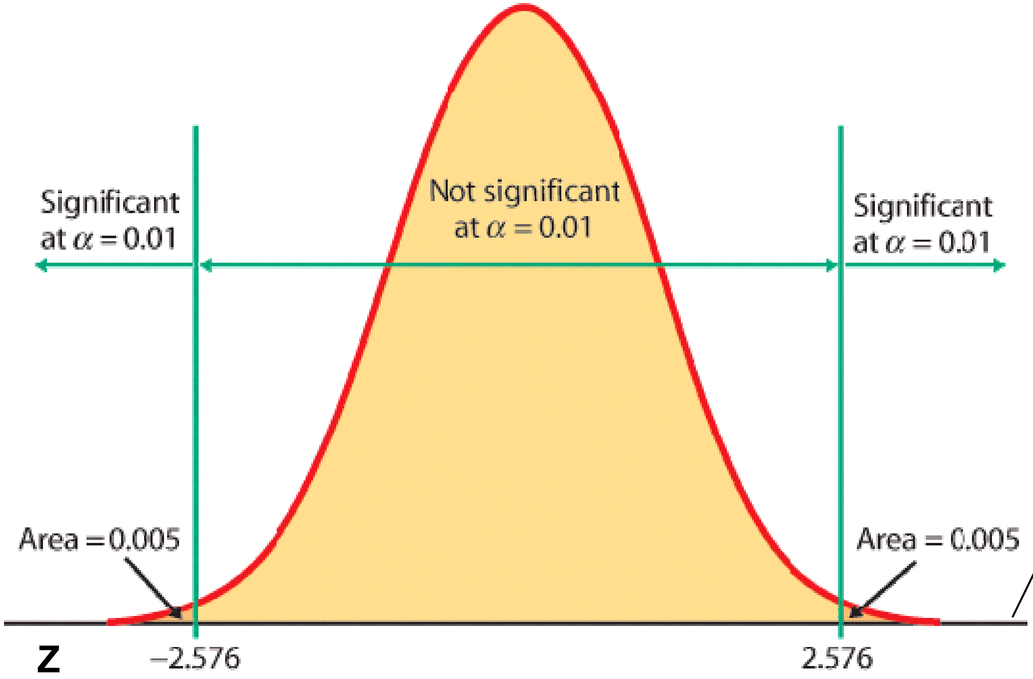
* If α had been set to 5%, then the P-value would be significant.

* If α had been set to 1%, then the P-value would <u>not</u> be significant.

When the z score falls within the rejection region (shaded area on the tail-side), the p-value is smaller than α and you have shown statistical significance.



One-sided test, α = 5%



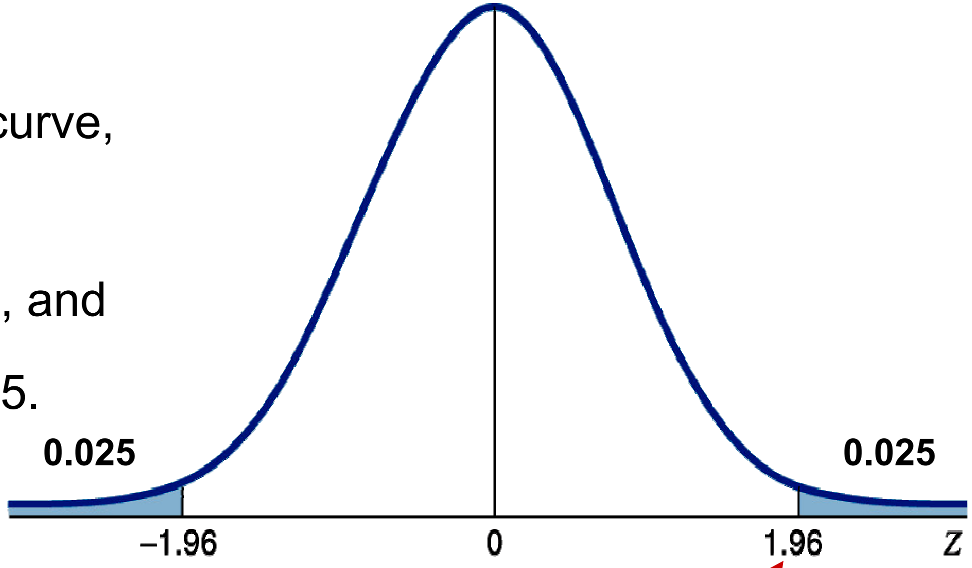
Two-sided test, α = 1%

Rejection region for a two-tail test of μ with $\alpha = 0.05$ (5%)

A two-sided test means that α is spread between both tails of the curve, thus:

-A middle area *C* of $1 - \alpha = 95\%$, and

-An upper tail area of $\alpha /2 = 0.025$.



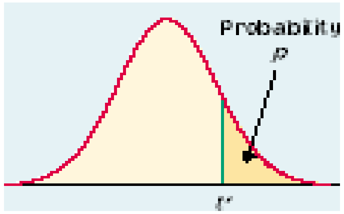


Table C

upper tail probability p	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
(…)												
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
Confidence interval C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Confidence intervals to test hypotheses

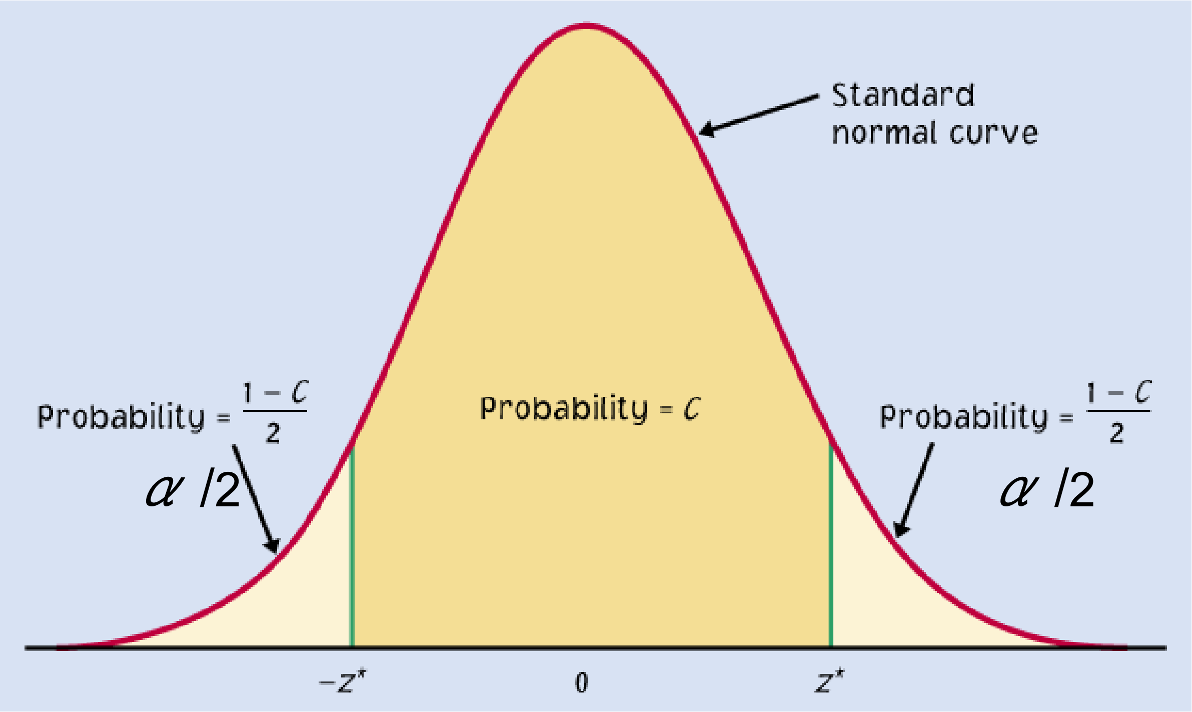
Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis.

In a two-sided test,

$C = 1 - \alpha$.

C confidence level

α significance level



Packs of cherry tomatoes (σ = 5 g): $H_0 : \mu$ = 227 g versus $H_a : \mu \neq$ 227 g

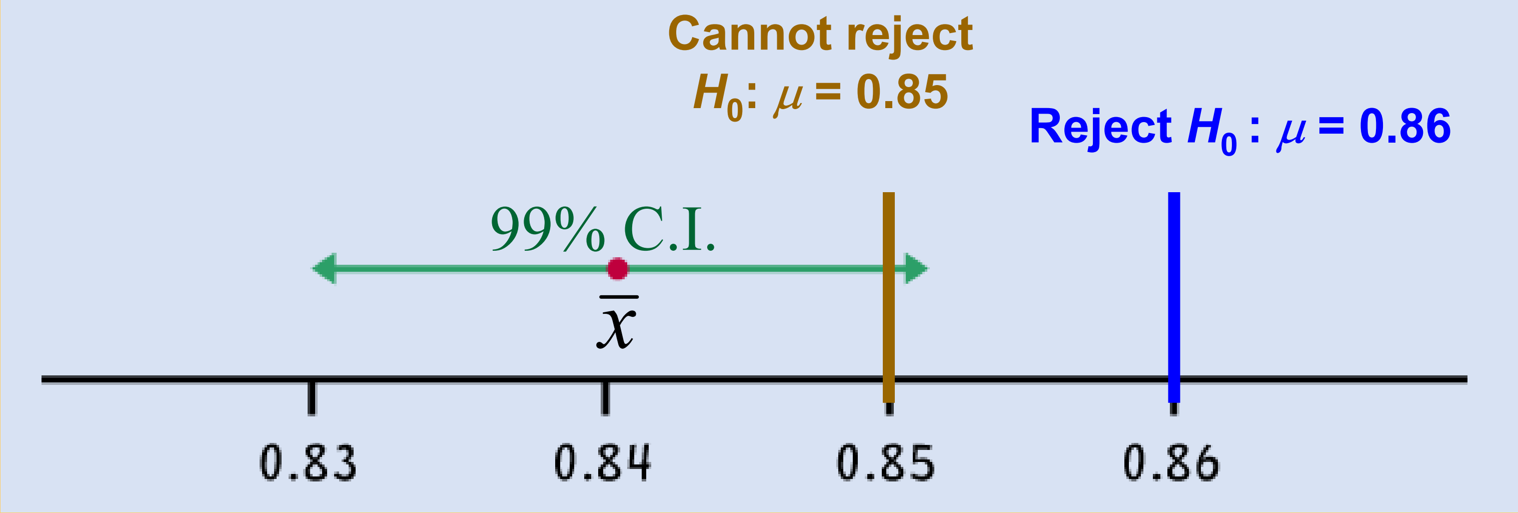
Sample average 222 g. 95% CI for μ = 222 ± 1.96*5/$\sqrt{4}$ = 222 g ± 4.9 g

227 g does not belong to the 95% CI (217.1 to 226.9 g). Thus, we reject H_0.

Logic of confidence interval test

Ex: Your sample gives a 99% confidence interval of $\bar{x} \pm m = 0.84 \pm 0.0101$.

With 99% confidence, could samples be from populations with μ = 0.86? μ = 0.85?



A confidence interval gives a black and white answer: Reject or don't reject H_0.

But it also estimates a range of likely values for the true population mean μ.

A P-value quantifies how strong the evidence is against the H_0. But if you reject H_0, it doesn’t provide any information about the true population mean μ.

Introduction to Inference

6.3 Use and Abuse of Tests
6.4 Power and Decision

© 2012 W.H. Freeman and Company

Objectives

6.3 Use and abuse of tests

6.4 Power and inference as a decision

- Cautions about significance tests
- Power of a test
- Type I and II errors
- Error probabilities

Cautions about significance tests

Choosing the significance level α

Factors often considered:

- What are the consequences of rejecting the null hypothesis (e.g., global warming, convicting a person for life with DNA evidence)?
- Are you conducting a preliminary study? If so, you may want a larger α so that you will be less likely to miss an interesting result.

Some conventions:

- We typically use the standards of our field of work.
- There are no “sharp” cutoffs: e.g., 4.9% versus 5.1 %.
- It is the order of magnitude of the P-value that matters: “somewhat significant,” “significant,” or “very significant.”

Practical significance

Statistical significance only says whether the effect observed is likely to be due to chance alone because of random sampling.

Statistical significance may not be practically important. That's because statistical significance doesn't tell you about the **magnitude** of the effect, only that there is one.

An effect could be too small to be relevant. And with a large enough sample size, significance can be reached even for the tiniest effect.

- A drug to lower temperature is found to reproducibly lower patient temperature by 0.4° Celsius (P-value < 0.01). But clinical benefits of temperature reduction only appear for a 1° decrease or larger.

Don’t ignore lack of significance

- Consider this provocative title from the British Medical Journal: “Absence of evidence is not evidence of absence.”
- Having no proof of who committed a murder does not imply that the murder was not committed.

Indeed, failing to find statistical significance in results is not rejecting the null hypothesis. This is very different from actually accepting it. The sample size, for instance, could be too small to overcome large variability in the population.

When comparing two populations, lack of significance does not imply that the two samples come from the same population. They could represent two very distinct populations with similar mathematical properties.

Interpreting effect size: It’s all about context

There is no consensus on how big an effect has to be in order to be considered meaningful. In some cases, effects that may appear to be trivial can be very important.

- Example: Improving the format of a computerized test reduces the average response time by about 2 seconds. Although this effect is small, it is important since this is done millions of times a year. The *cumulative* time savings of using the better format is gigantic.

Always think about the context. Try to plot your results, and compare them with a baseline or results from similar studies.

The power of a test

The **power** of a test of hypothesis with fixed significance level α is the probability that the test will reject the null hypothesis when the alternative is true.

In other words, power is the probability that the data gathered in an experiment will be sufficient to reject a wrong null hypothesis.

Knowing the power of your test is important:

- When designing your experiment: select a sample size large enough to detect an effect of a magnitude you think is meaningful.
- When a test found no significance: Check that your test would have had enough power to detect an effect of a magnitude you think is meaningful.

Test of hypothesis at significance level α 5%:

H_0: $\mu = 0$ versus H_a: $\mu > 0$

Can an exercise program increase bone density? From previous studies, we assume that $\sigma = 2$ for the percent change in bone density and would consider a percent increase of 1 medically important.
Is 25 subjects a large enough sample for this project?

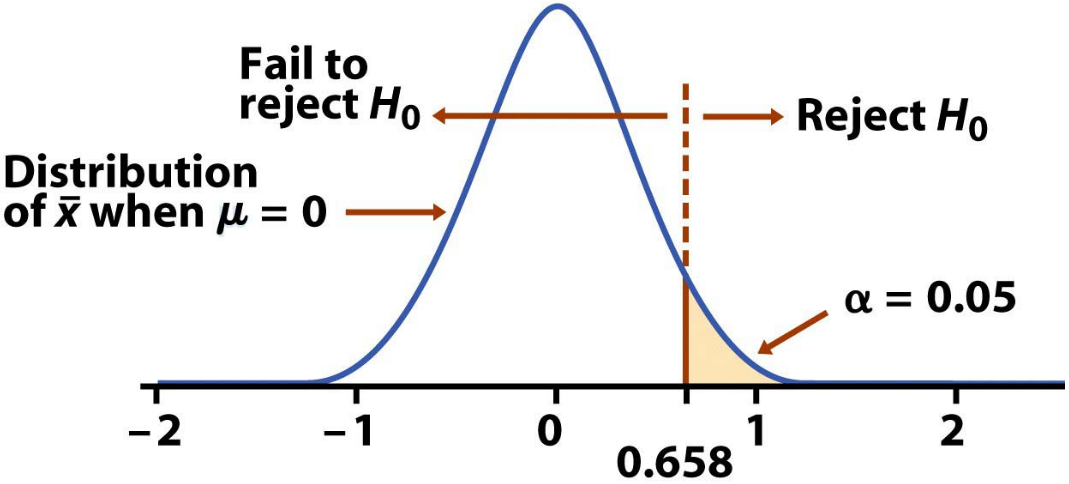
A significance level of 5% implies a lower tail of 95% and $z = 1.645$. Thus:

$$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$$

$$\bar{x} = \mu + z * (\sigma / \sqrt{n})$$

$$\bar{x} = 0 + 1.645 * (2 / \sqrt{25})$$

$$\bar{x} = 0.658$$



All sample averages larger than 0.658 will result in rejecting the null hypothesis.

What if the null hypothesis is wrong and the true population mean is 1?

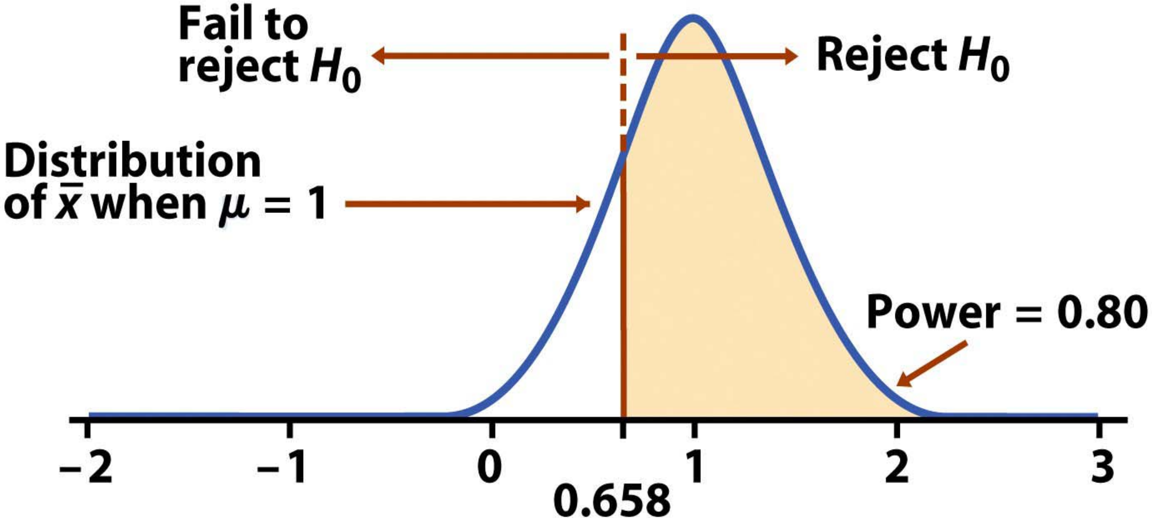
The **power against the alternative** μ = 1% is the probability that H_0 will be rejected when in fact μ = 1%.

$$= P(\bar{x} \geq 0.658 \text{ when } \mu = 1)$$

$$= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.658 - 1}{2/\sqrt{25}}\right)$$

$$= P(z > -0.855) = 0.80$$

We expect that a sample size of 25 would yield a power of 80%.



A test power of 80% or more is considered good statistical practice.

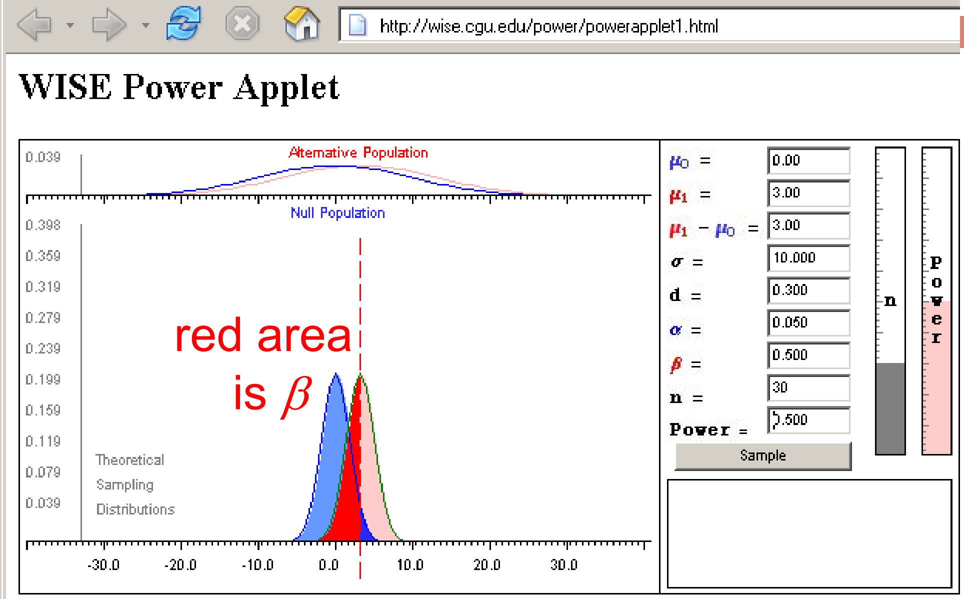
Factors affecting power: Size of the effect

The **size of the effect** is an important factor in determining power. Larger effects are easier to detect.

More conservative **significance levels** (lower α) yield lower power. Thus, using an α of .01 will result in less power than using an α of .05.

Increasing the **sample size** decreases the spread of the sampling distribution and therefore increases power. But there is a tradeoff between gain in power and the time and cost of testing a larger sample.

A larger **variance** σ^2 implies a larger spread of the sampling distribution, σ/sqrt(N). Thus, the larger the variance, the lower the power. The variance is in part a property of the population, but it is possible to reduce it to some extent by carefully designing your study.



http://wise.cgu.edu/power/powerapplet1.html

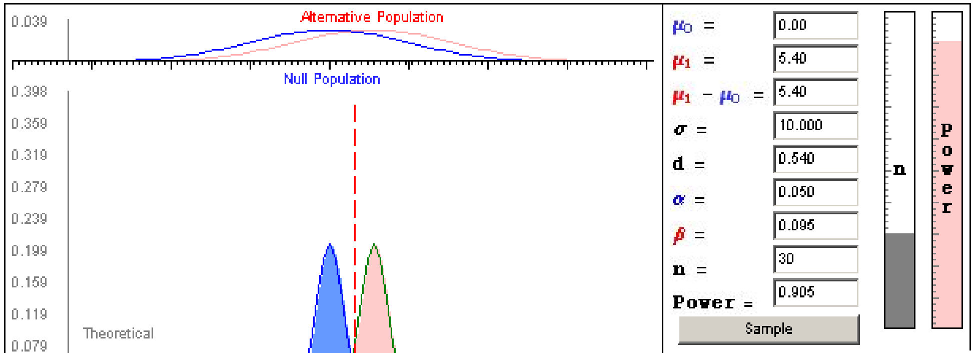
WISE Power Applet

μ_0 = 0.00
μ_1 = 3.00
$\mu_1 - \mu_0$ = 3.00
σ = 10.000
d = 0.300
α = 0.050
β = 0.500
n = 30
Power = .500

red area is β

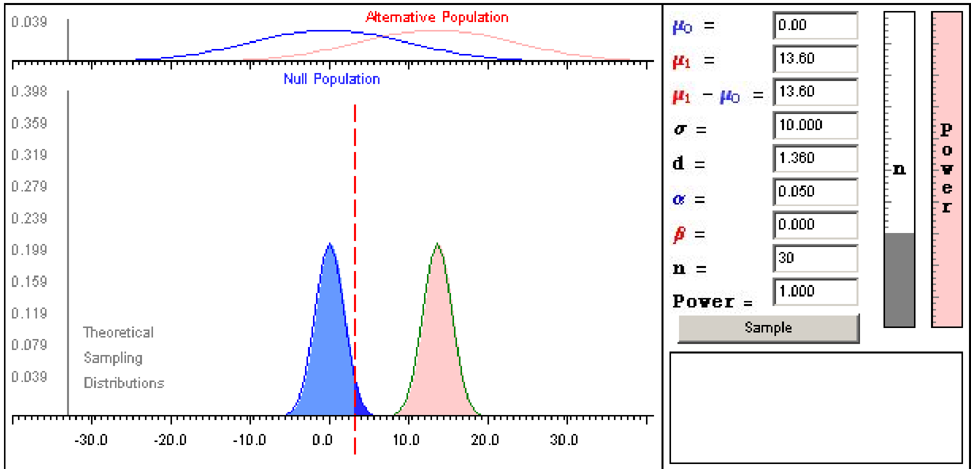
H_o: $\mu = 0$
$\sigma = 10$
$n = 30$
$\alpha = 5\%$

1. Real μ is 3 => power = .5
2. Real μ is 5.4 => power = .905
3. Real μ is 13.5 => power = 1



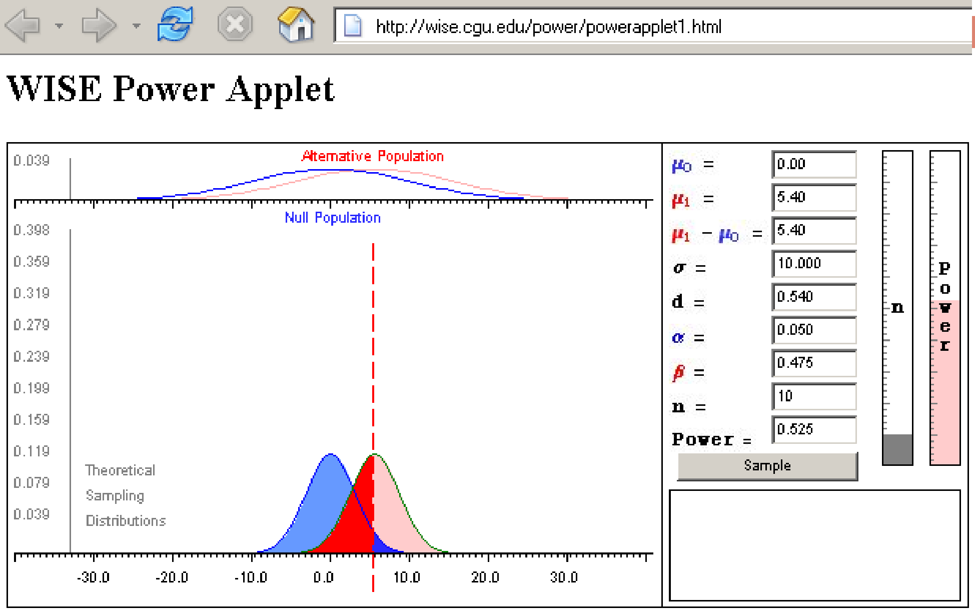
μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 10.000
d = 0.540
α = 0.050
β = 0.095
n = 30
Power = 0.905

→ larger differences are easier to detect



μ_0 = 0.00
μ_1 = 13.60
$\mu_1 - \mu_0$ = 13.60
σ = 10.000
d = 1.360
α = 0.050
β = 0.000
n = 30
Power = 1.000

http://wise.cgu.edu/power/power_applet.html

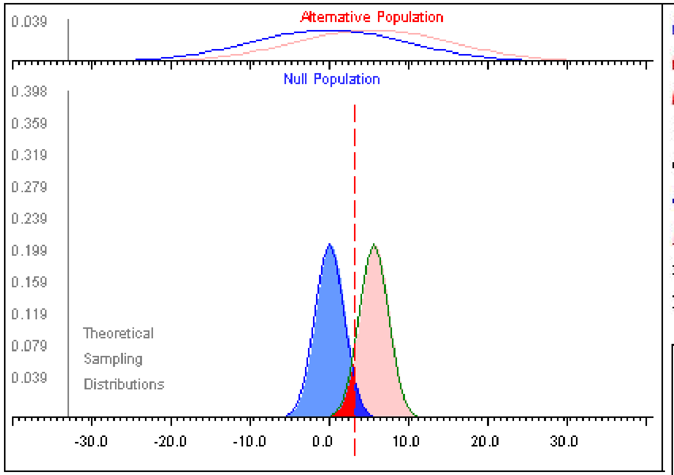


WISE Power Applet

μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 10.000
d = 0.540
α = 0.050
β = 0.475
n = 10
Power = 0.525

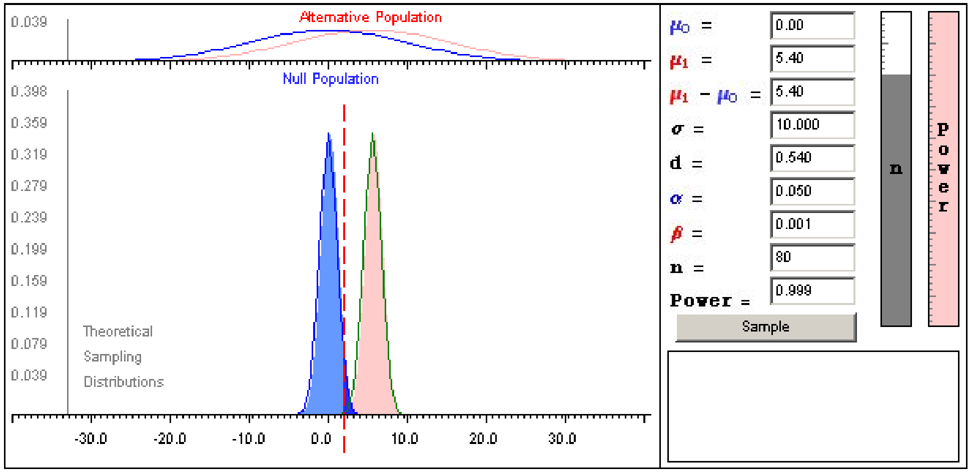
H_o: $\mu = 0$
$\sigma = 10$
Real $\mu = 5.4$
$\alpha = 5\%$

1. $n = 10$ => power = .525
2. $n = 30$ => power = .905
3. $n = 80$ => power = .999



μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 10.000
d = 0.540
α = 0.050
β = 0.095
n = 30
Power = 0.905
Sample

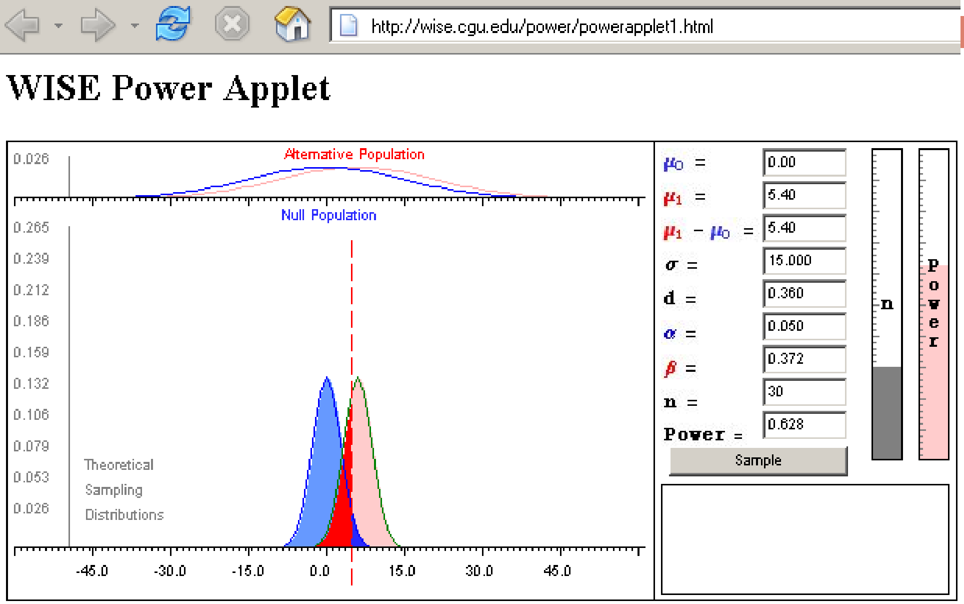
➔ larger sample sizes yield greater power



μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 10.000
d = 0.540
α = 0.050
β = 0.001
n = 80
Power = 0.999

*H*o: *μ* = 0
Real *μ* = 5.4
n = 30
α = 5%

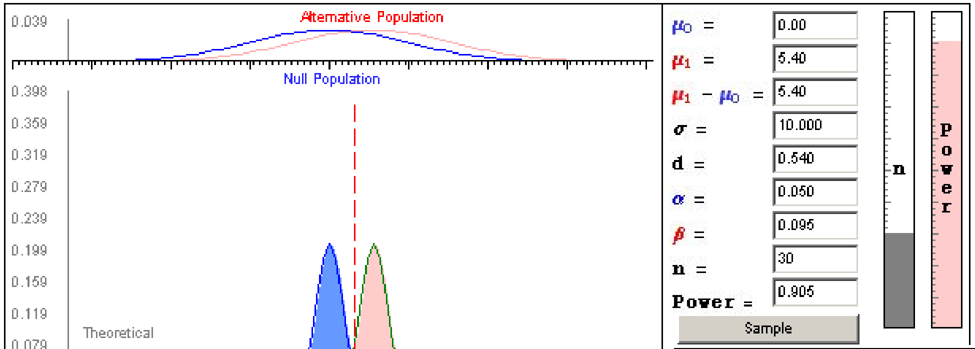
1. *σ* is 15 => power = .628
2. *σ* is 10 => power = .905
3. *σ* is 5 => power = 1



http://wise.cgu.edu/power/powerapplet1.html

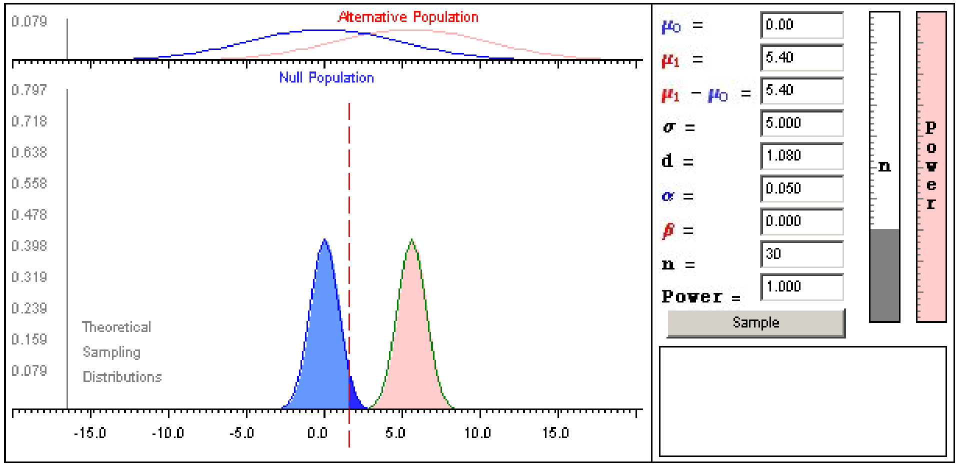
WISE Power Applet

μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 15.000
d = 0.360
α = 0.050
β = 0.372
n = 30
Power = 0.628



μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 10.000
d = 0.540
α = 0.050
β = 0.095
n = 30
Power = 0.905

➔ smaller variability yields greater power



μ_0 = 0.00
μ_1 = 5.40
$\mu_1 - \mu_0$ = 5.40
σ = 5.000
d = 1.080
α = 0.050
β = 0.000
n = 30
Power = 1.000

Type I and II errors

- A **Type I error** is made when we reject the null hypothesis and the null hypothesis is actually true (incorrectly reject a true H_0).

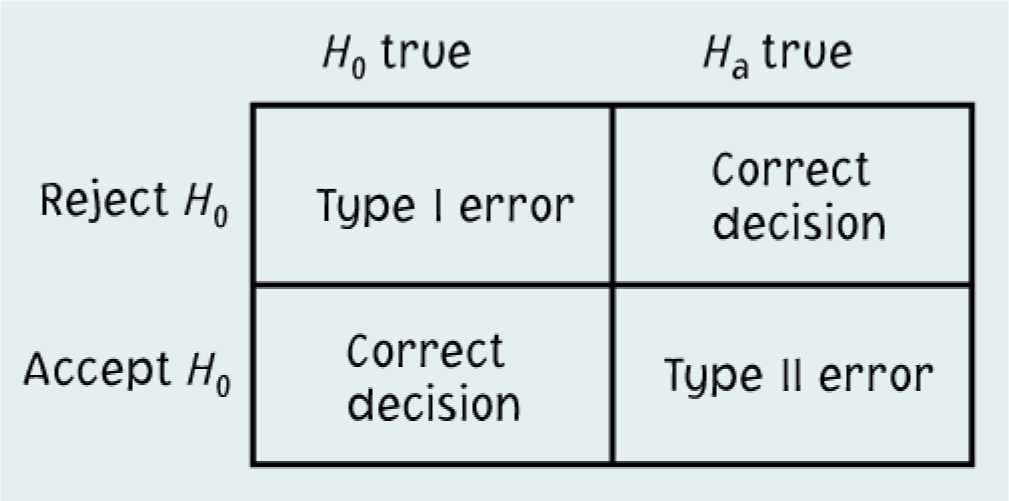
  The probability of making a Type I error is the significance level α.

- A **Type II error** is made when we fail to reject the null hypothesis and the null hypothesis is false (incorrectly keep a false H_0).

  The probability of making a Type II error is labeled β.

  The power of a test is $1 - \beta$.

Running a test of significance is a balancing act between the chance α of making a **Type I error** and the chance β of making a **Type II error.** Reducing α reduces the power of a test and thus increases β.



	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

It might be tempting to emphasize greater power (the more the better).

- However, with “too much power” trivial effects become highly significant.
- A type II error is not definitive since a failure to reject the null hypothesis does not imply that the null hypothesis is wrong.

The Common Practice of Testing of Hypotheses

1. State H_o and H_a as in a test of significance.

2. Think of the problem as a decision problem, so the probabilities of Type I and Type II errors are relevant.

3. Consider only tests in which the probability of a Type I error is no greater than α.

4. Among these tests, select a test that makes the probability of a Type II error as small as possible.

Alternate Slide

The following slide offers alternate software output data and examples for this presentation.

Steps for Tests of Significance

1. Assumptions/Conditions
   - Specify variable, parameter, method of data collection, shape of population.
2. State hypotheses
   - *Null hypothesis H_o* and *alternative hypothesis H_a*.
3. Calculate value of the test statistic
   - A measure of "difference" between hypothesized value and its estimate.
4. Determine the *P-value*
   - Probability, assuming H_o true that the test statistic takes the observed value or a more extreme value.
5. State the decision and conclusion
   - Interpret P-value, make decision about H_o.