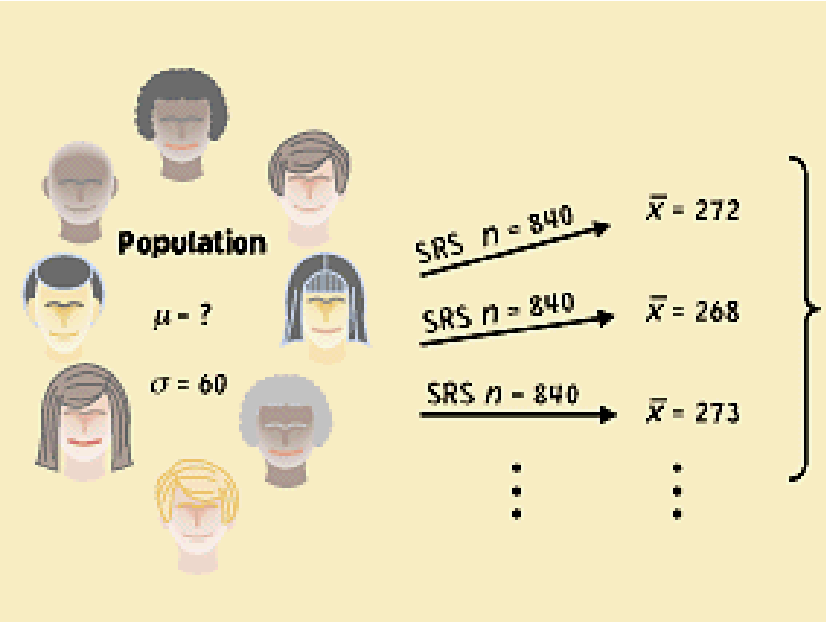
Lecture 4

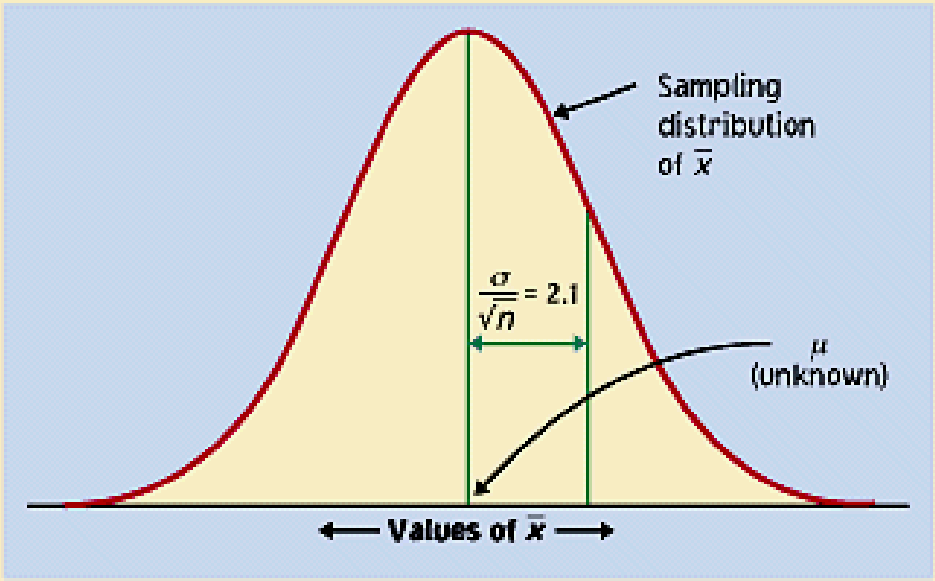
Statistical Inference. Inference for one population mean and one population proportion

Uncertainty and confidence

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

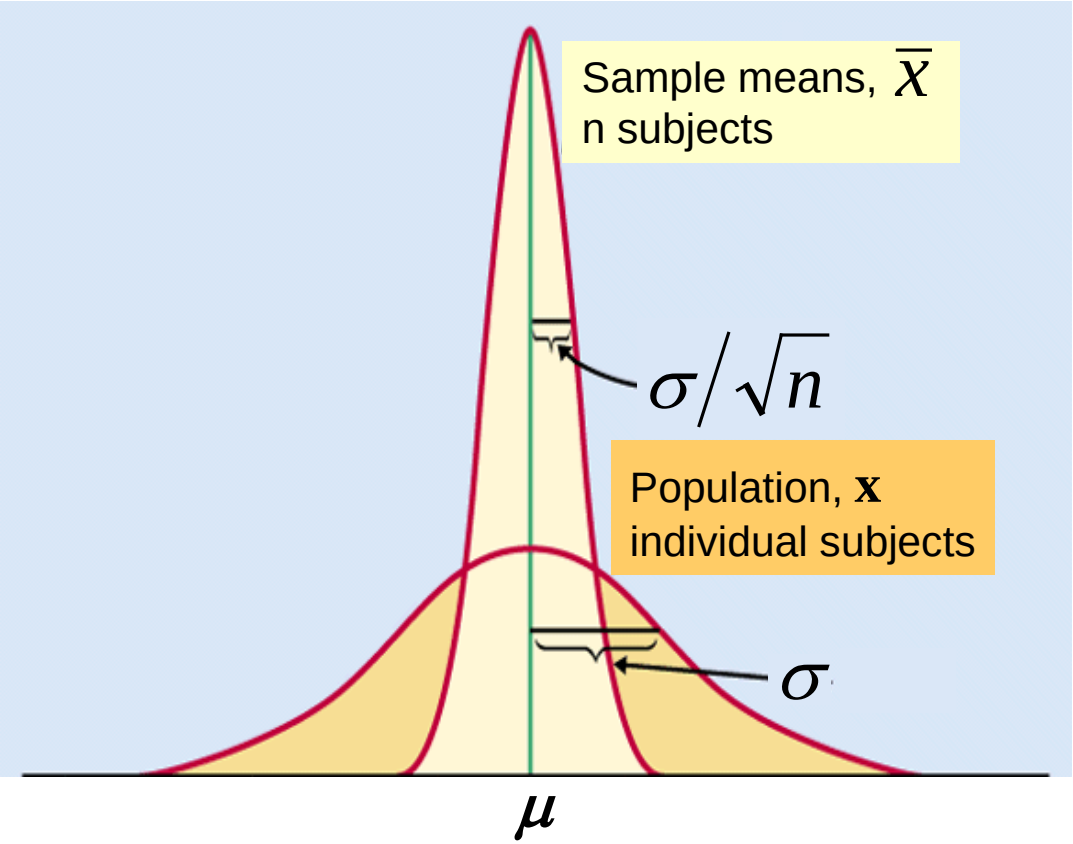
In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ.





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

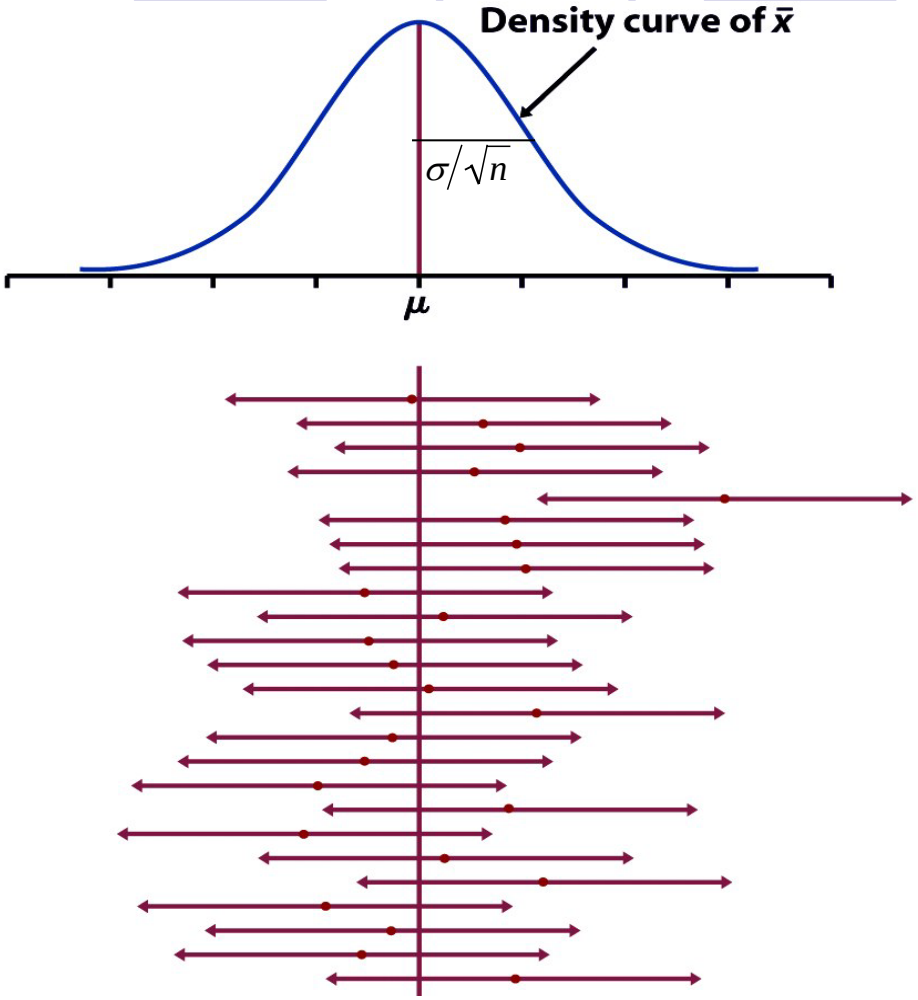
Thus, the estimates $\bar{x}$ gained from our samples are always relatively close to the population parameter μ.



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.

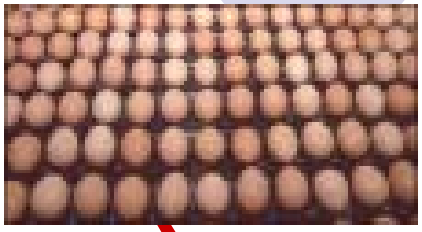
Because distances are symmetrical, this implies that **the population parameter μ must be within roughly 2 standard deviations from the sample average $\bar{x}$, in 95% of all samples.**



Red dot: mean value of individual sample

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.



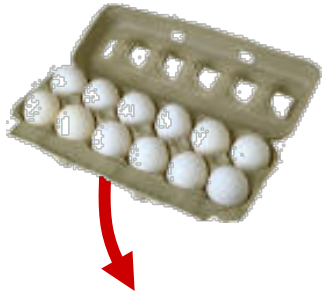
- What is the distribution of the sample means $\bar{x}$?

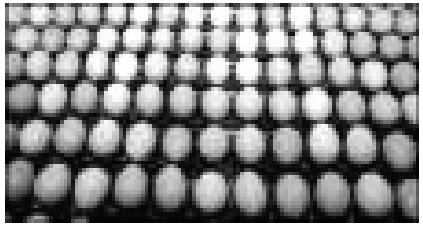
  Normal (mean μ, standard deviation $\sigma/\sqrt{n}$) = *N*(65 g,1.44 g).

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

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

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.



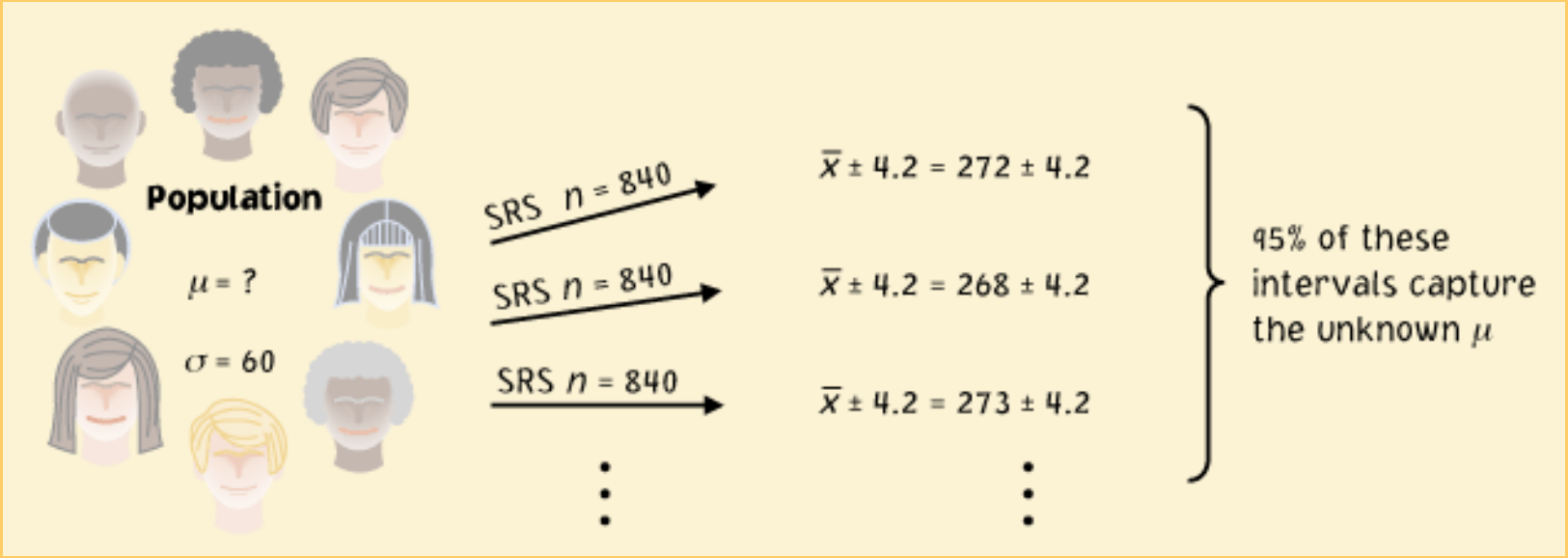


- Knowing that the standard deviation of egg weight is 5 g, what can you infer about the mean μ of the white egg population?

  There is a 95% chance that the population mean μ is roughly within $\pm 2\sigma/\sqrt{n}$ of $\bar{x}$, or 64.2 g ± 2.88 g.

Confidence interval

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.

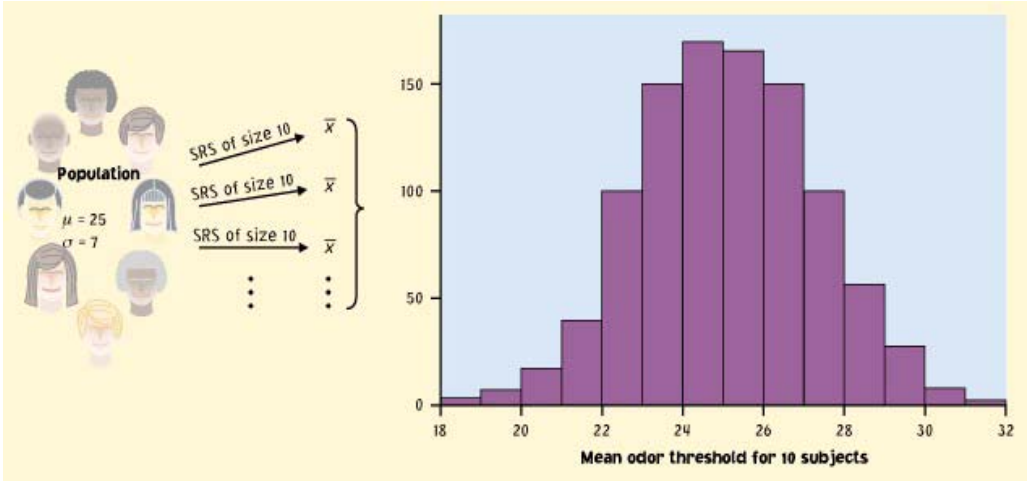


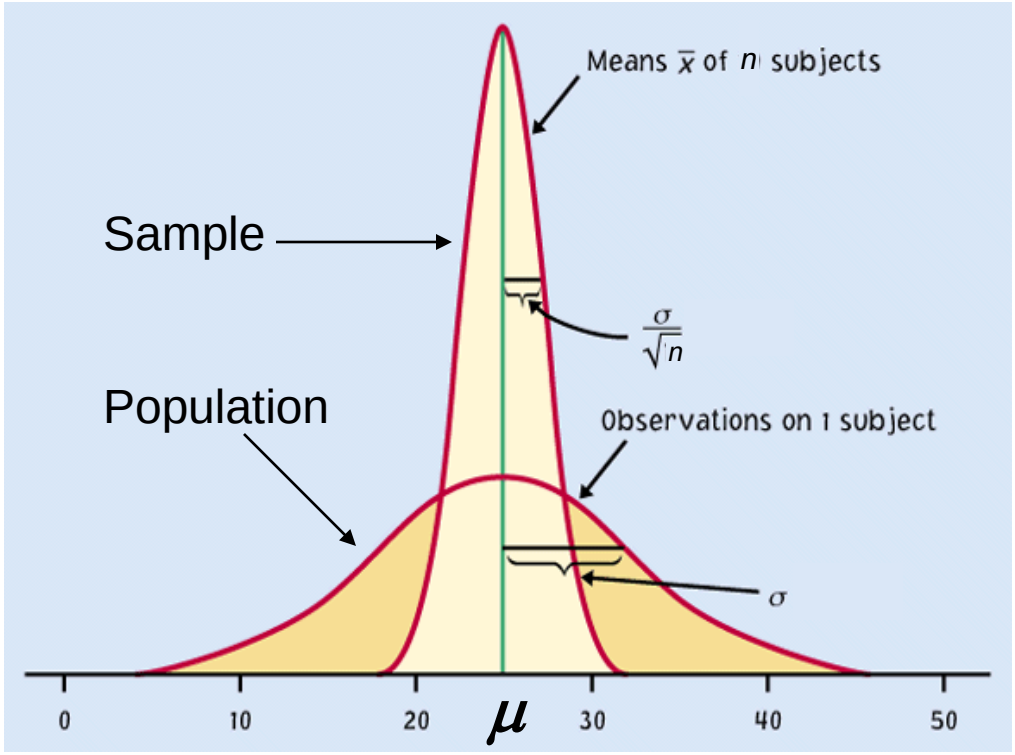
$\bar{x} \pm 4.2$ is a 95% confidence interval for the population parameter μ.

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.



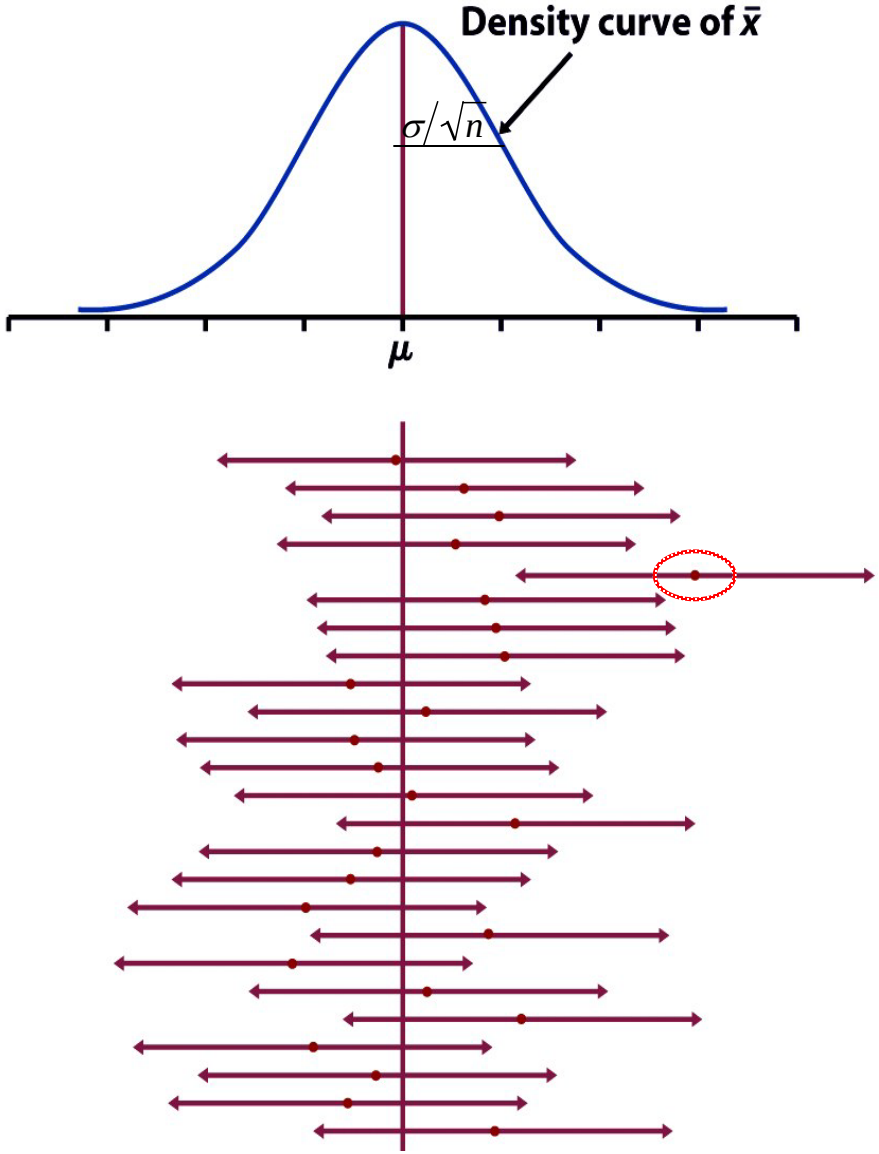


All we need is one SRS of size *n* and relying on the properties of the sample means distribution to infer the population mean μ.

Reworded

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

- 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.



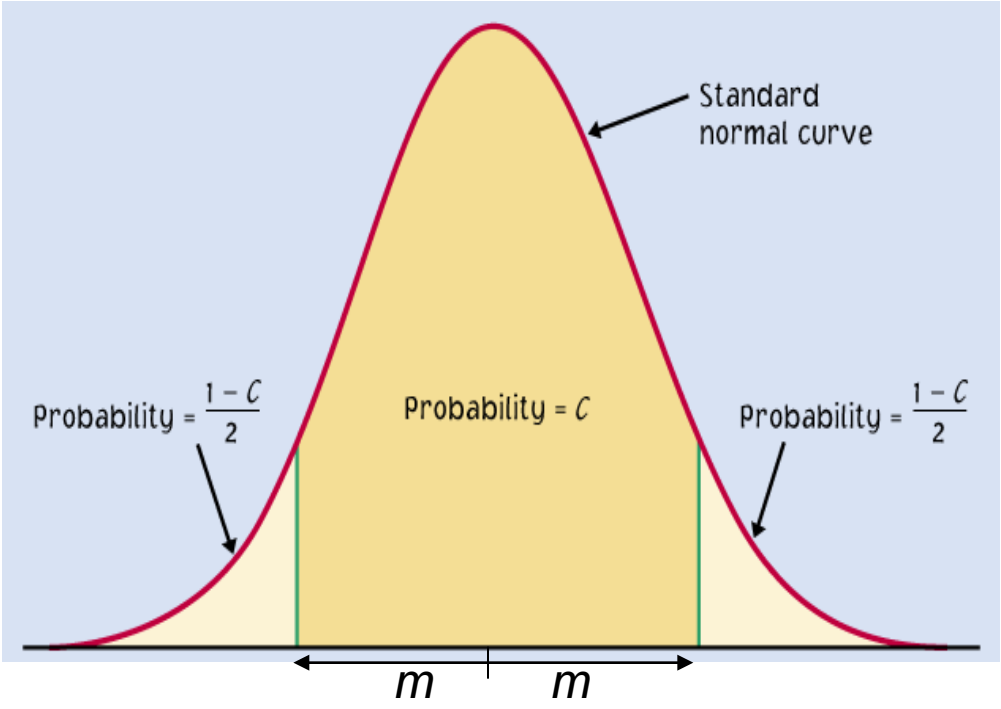
A **confidence interval** can be expressed as:

- Mean ± *m*
  m is called the **margin of error**
  μ within $\overline{x} \pm m$
  Example: 120 ± 6

- Two endpoints of an interval
  μ within ($\overline{x} - m$) to ($\overline{x} + m$)
  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.



Varying confidence levels

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

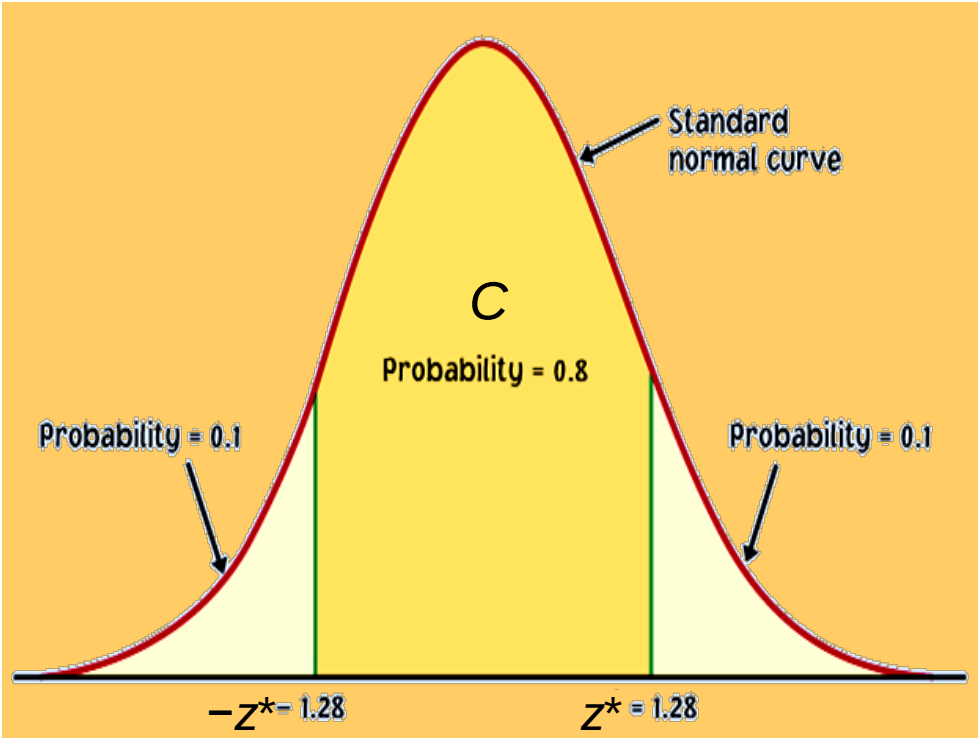
Different areas under the curve give different confidence levels *C*.

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. 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 **R:**

qnorm(probability,mean,standard_dev)

gives z quantile for a given probability.

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

Example: For a 98% confidence level, qnorm(0.01,0,1) = −2.326348 (= neg. *z**)

Link between confidence level and margin of error

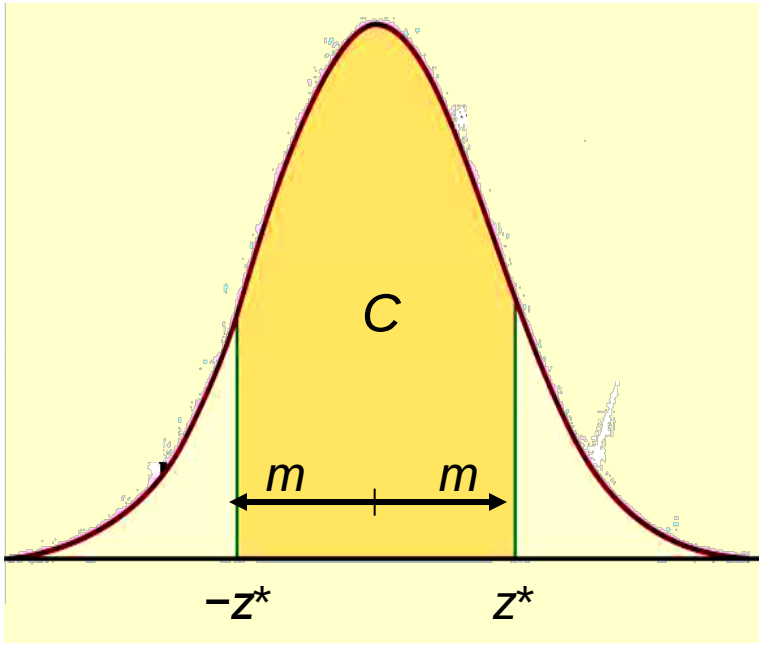
The confidence level *C* determines the value of *z**

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

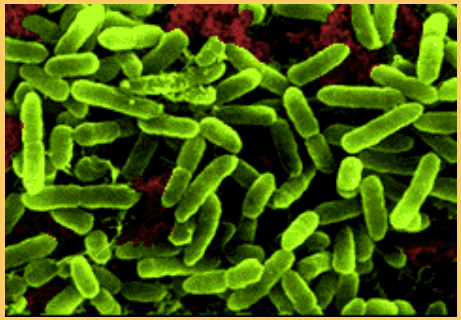
$$m = z^* \sigma / \sqrt{n}$$

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

A lower 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

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

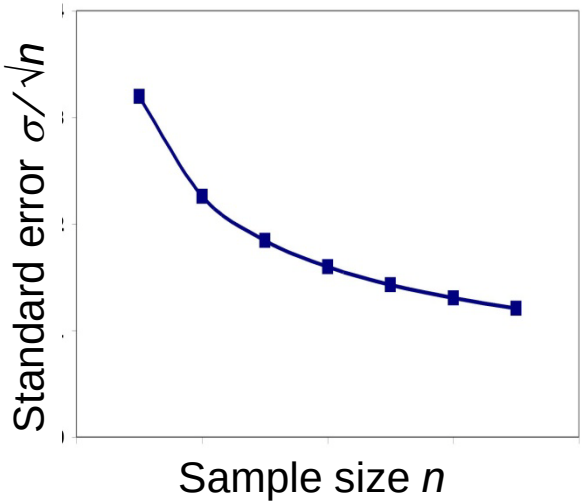
$= 28 \pm 0.60 \times 10^6$ bacteria/ml

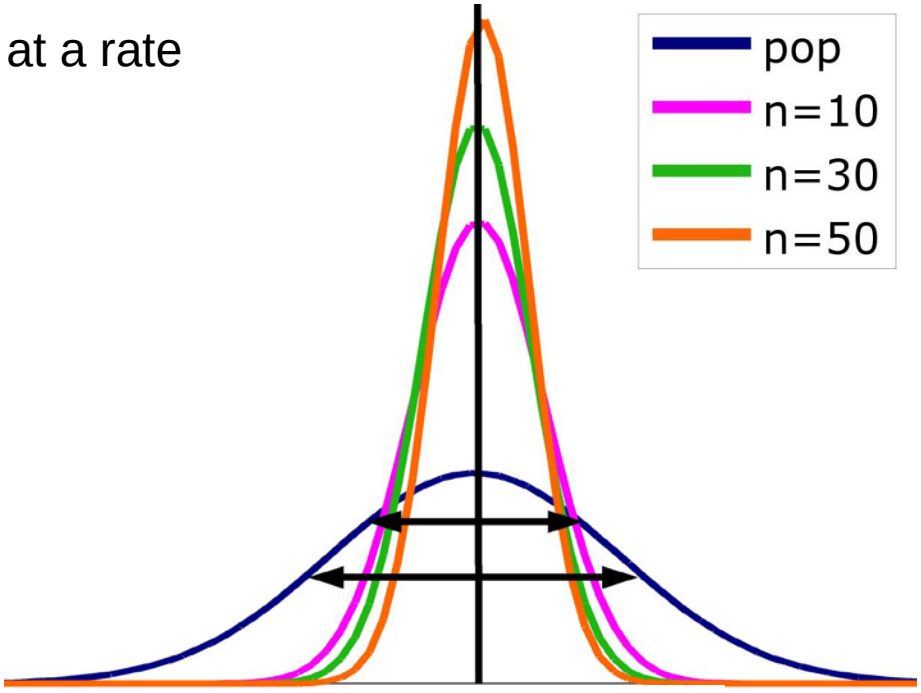
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

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

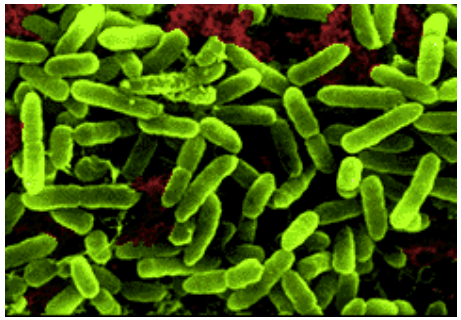
You may need a certain margin of error (e.g., drug trial, manufacturing specs). In many cases, the population variability (σ) is fixed, but we can choose the number of measurements (*n*).

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 $\sigma = 1 * 10^6$ 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 90%?

For a 90% confidence interval, $z^* = 1.645$.

$$n = \left(\frac{z^*\sigma}{m}\right)^2 \Rightarrow n = \left(\frac{1.645*1}{0.5}\right)^2 = 3.29^2 = 10.8241$$

Using only 10 measurements will not be enough to ensure that m is no more than 0.5 * 106. Therefore, we need at least 11 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:

- Data: a SRS
- Formulas for other randomized designs available
- Haphazard data = unreliable conf. int.
- Population need not be normal (our Example 1 wasn't) but outliers pose a threat to validity of conclusions.
- We need to know σ the population variability. We will recall how to deal with the usually unknown σ later in this lecture.

Section: Tests of Significance

- The scheme of reasoning
- Stating hypotheses
- Test statistics
- P-values
- Statistical significance
- Test for population mean
- Two-sided test and confidence intervals

We have seen that the properties of the sampling distribution of *x* bar 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?

Null and alternative 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 : μ = 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 : μ = [a specific number] H_a : $\mu \neq$ [a specific number]

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

H_0 : μ = [a specific number] H_a : $\mu <$ [a specific number] OR

H_0 : μ = [a specific number] H_a : $\mu >$ [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:

$H_0 : \mu_{\text{generic}} = \mu_{\text{brand}}$ $H_a : \mu_{\text{generic}} \neq \mu_{\text{brand}}$ 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 in 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 because of 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.

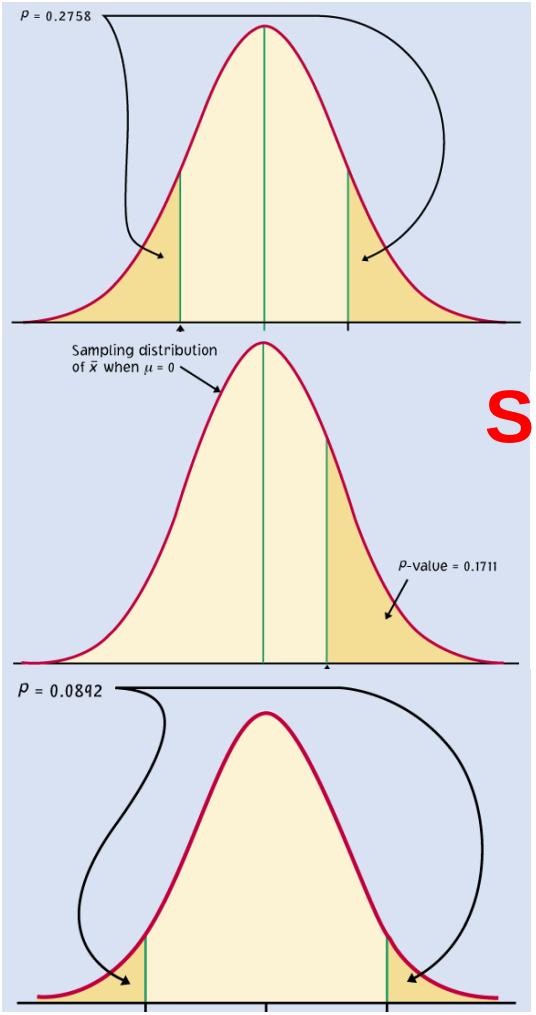
Thus, small P-values are strong evidence AGAINST H_0.

But how small is small…?

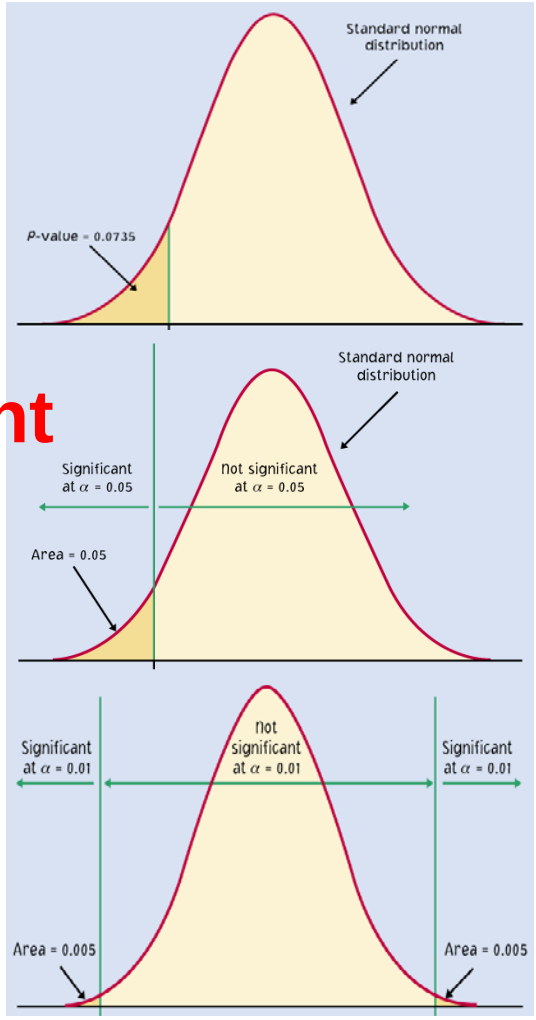
P = 0.2758

P = 0.1711

P = 0.0892



Significant P-value ???



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.

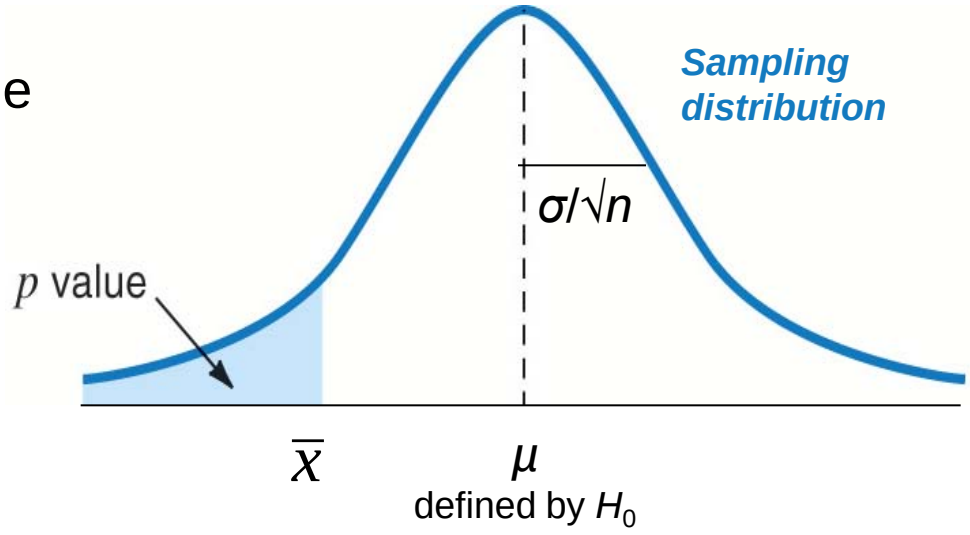
Tests for a population mean

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.

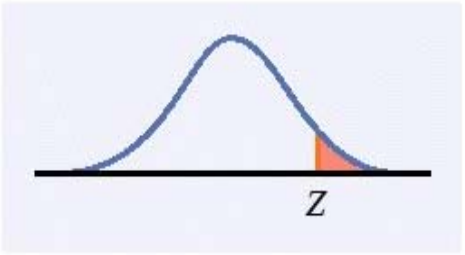
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



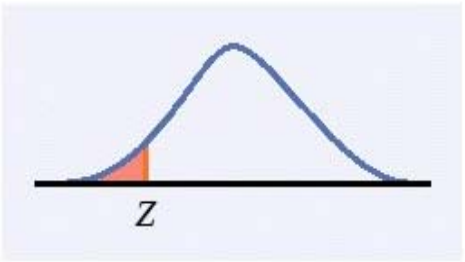
P-value in one-sided and two-sided tests

One-sided (one-tailed) test

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

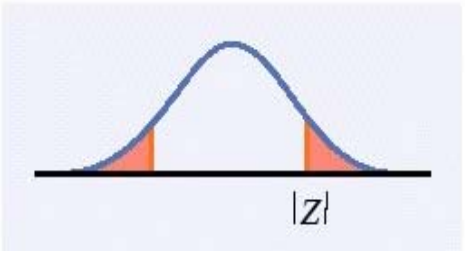


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



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 : μ = 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?

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

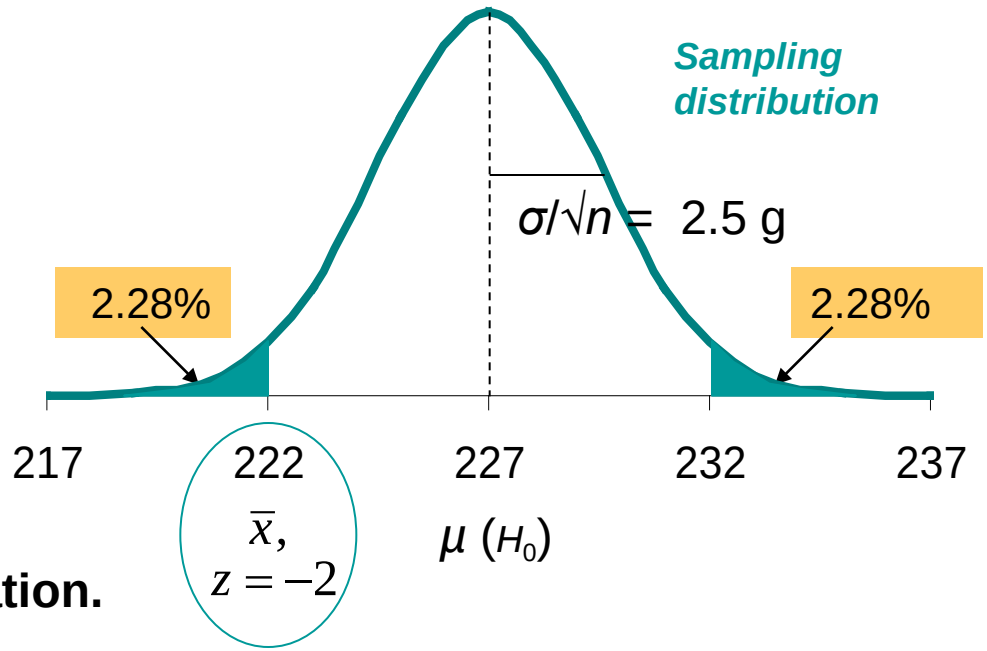
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{4}} = -2$$

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.**



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 before conducting the test.

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

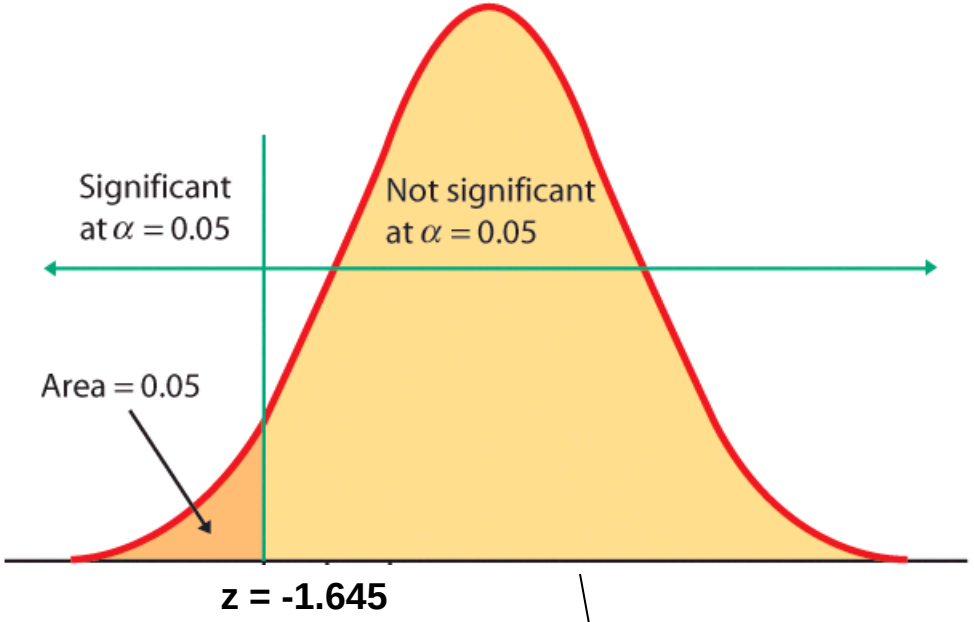
Does the packaging machine need revision?

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

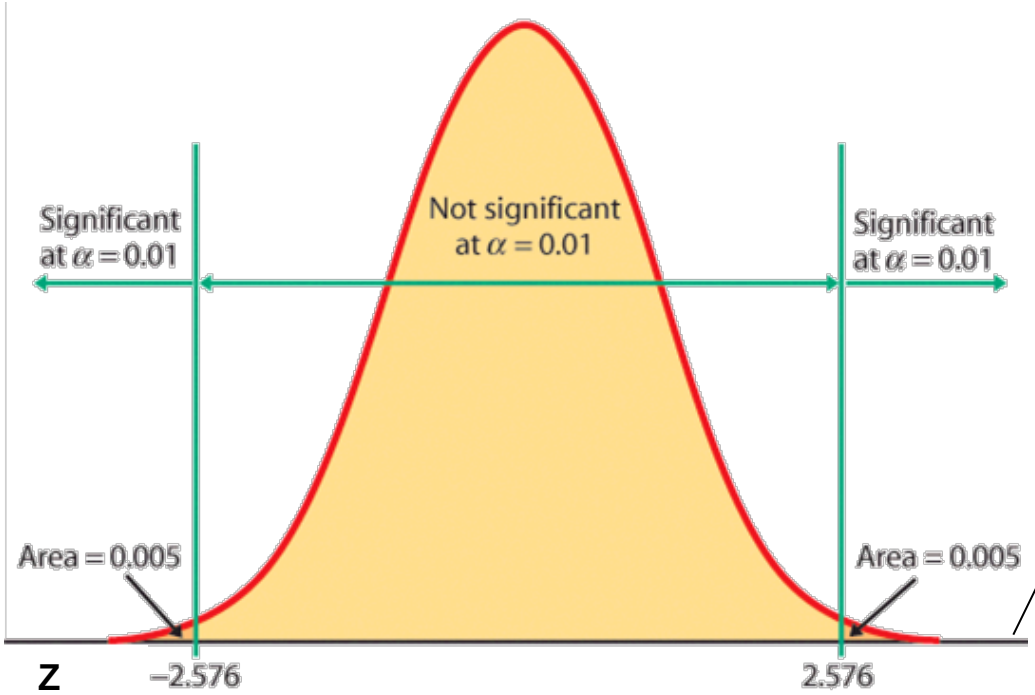
* If α had been set to 5%, then the P-value would be significant.

* If α had been set to 1%, then the P-value would not 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%



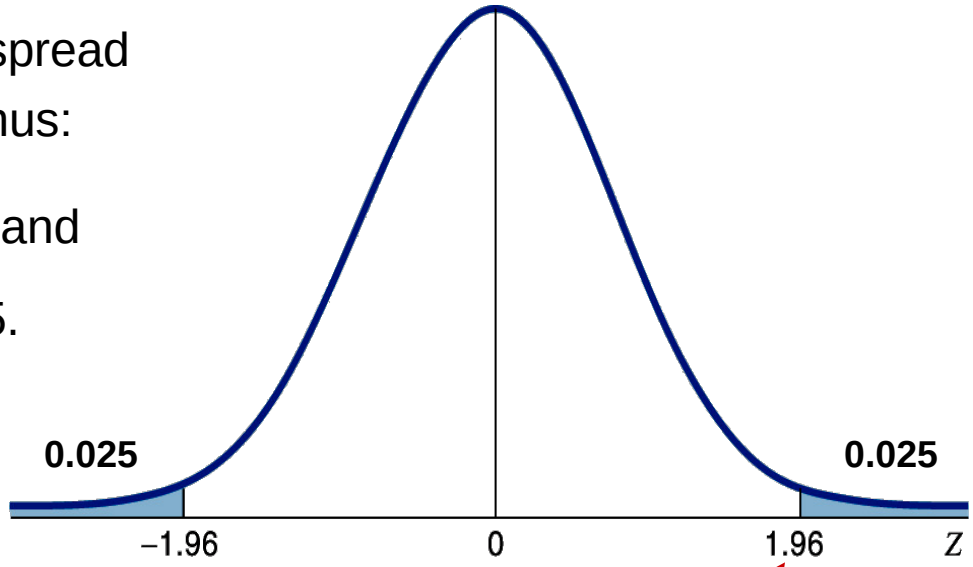
Two-sided test, α = 1%

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 $\alpha/2 = 0.025$.



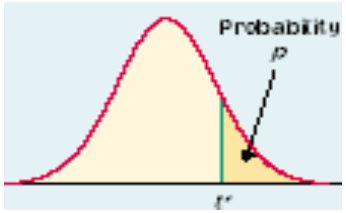


Table C

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

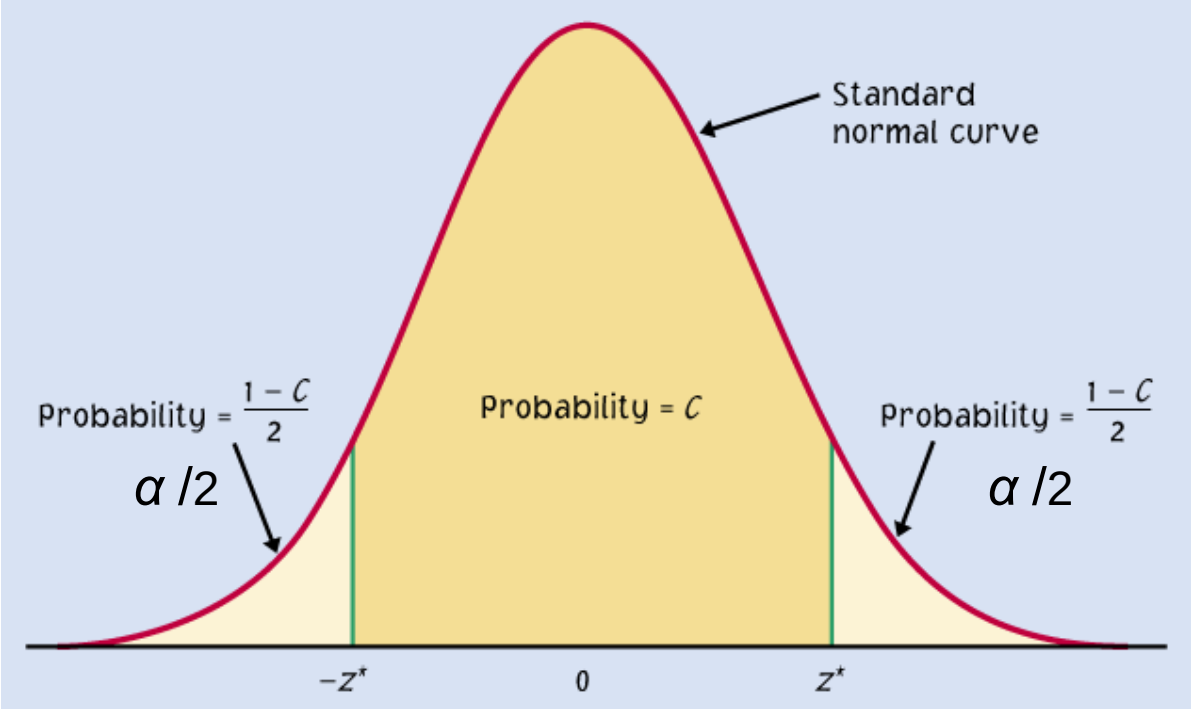
Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis.

In a two-sided test,

$C = 1 - \alpha$.

C confidence level

α significance level



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

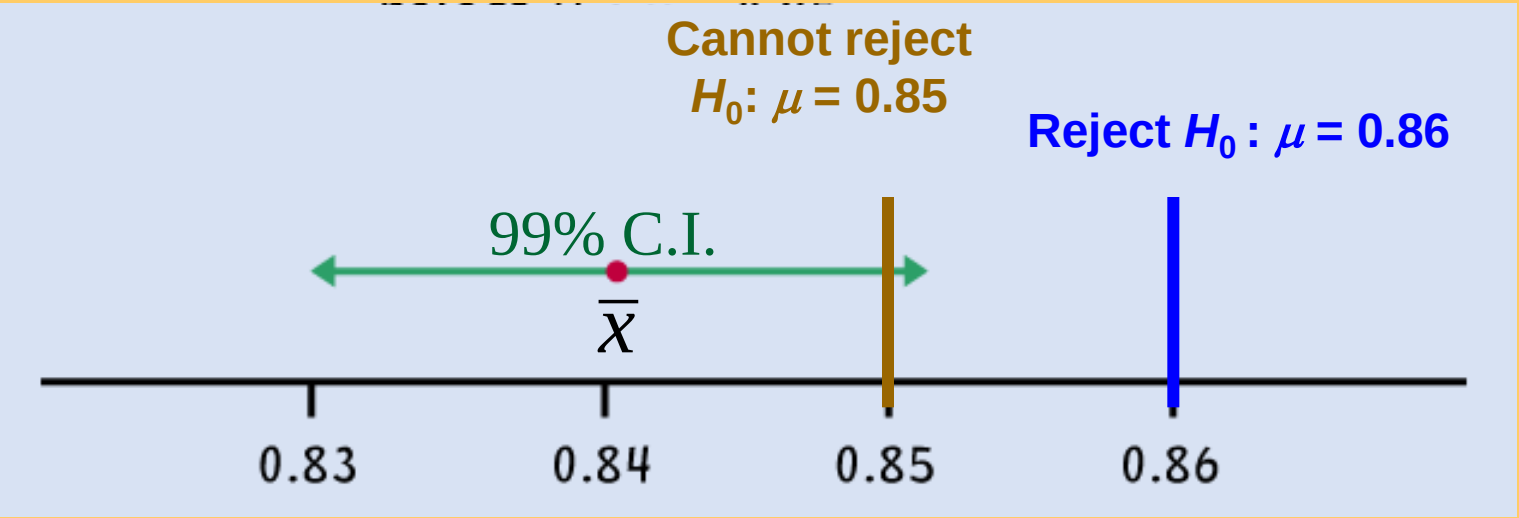
Sample average 222 g. 95% CI for μ = 222 ± 1.96*5/√4 = 222 g ± 4.9 g

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 $\mu = 0.86$? $\mu = 0.85$?



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.

Section: Use and abuse of tests

- Choosing the level of significance
- Significance vs. practical significance
- Lack of significance may be informative
- Dangers of searching for significance
- Assumptions about the data

Caution about significance tests

Choosing the significance level *α*

- α=0.05 is accepted standard, but…
  - if the conclusion that H_a is true has "costly" implications, smaller α may be appropriate

  e.g.,

  - 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 %. Oftentimes, describing the evidence using the P-value itself may be enough
  - 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.

Interpreting lack of significance

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

Indeed, failing to find statistical significance in results means that we do not reject 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 the 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 in reality 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: to 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?

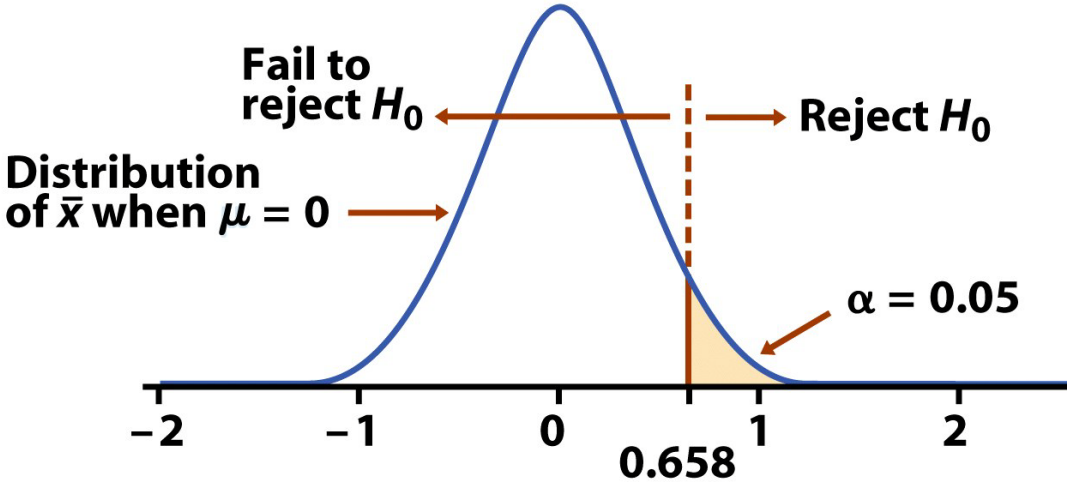
A significance level of 5% implies a lower tail of 95% and $z = 1.645$. Thus:

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

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

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

$$\bar{x} = 0.658$$



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?

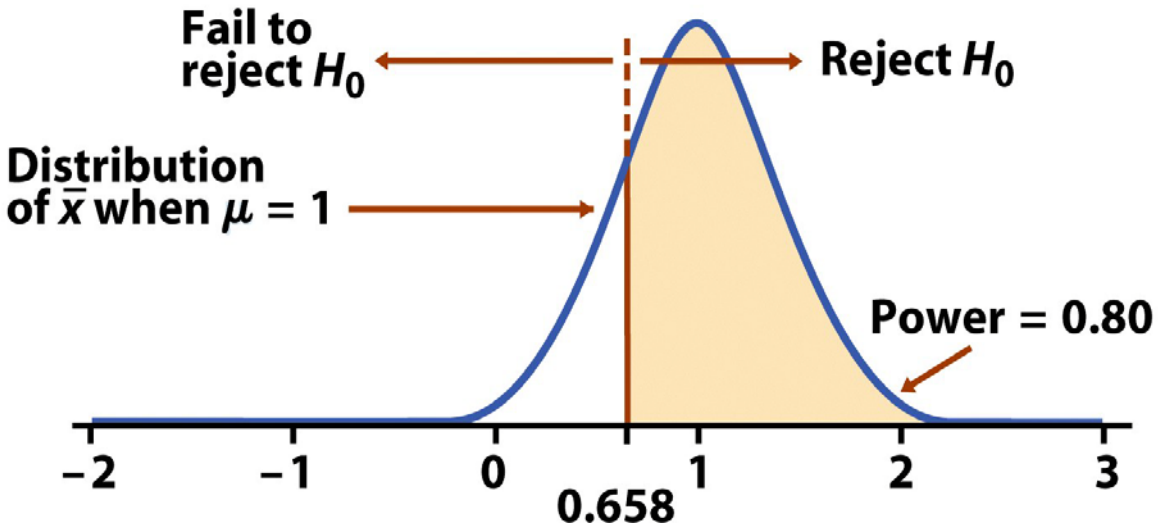
The **power against the alternative** $\mu = 1\%$ is the probability that H_0 will be rejected when in fact $\mu = 1\%$.

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

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

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

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.

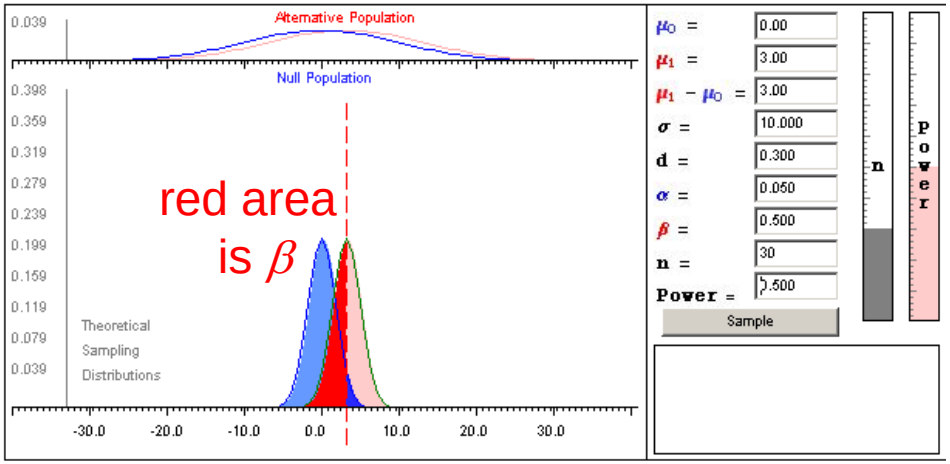
More conservative **significance levels** (lower α) yield lower power. Thus, using an α of .01 will result in lower power than using an α of .05.

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.

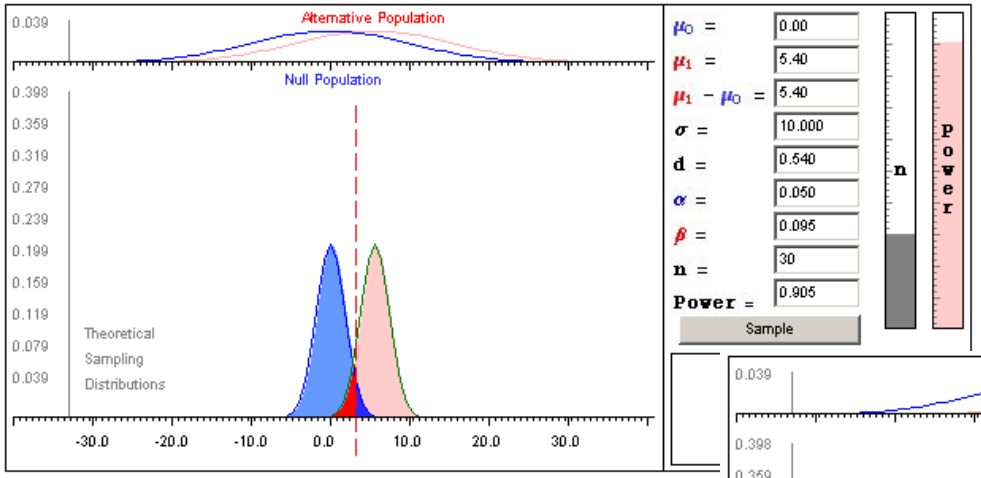
http://wise.cgu.edu/power/powerapplet1.html

WISE Power Applet

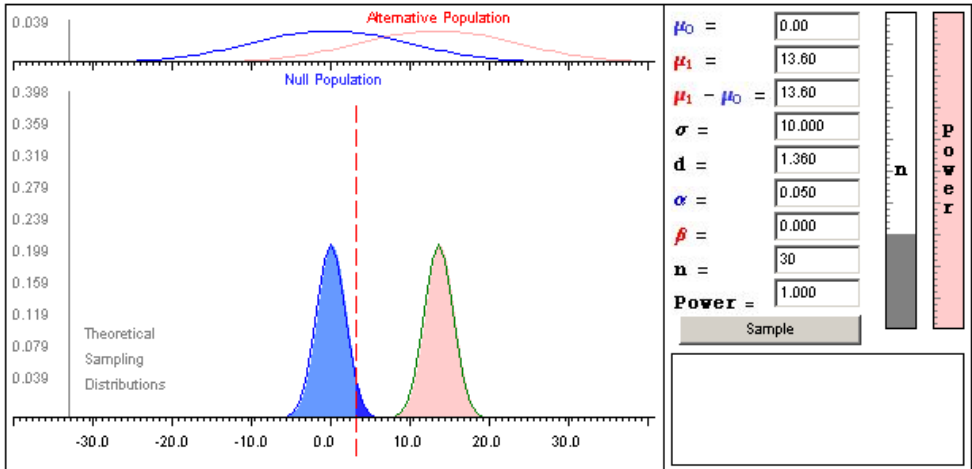


H_o: $\mu = 0$
$\sigma = 10$
$n = 30$
$\alpha = 5\%$

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



➔ **larger differences are easier to detect**



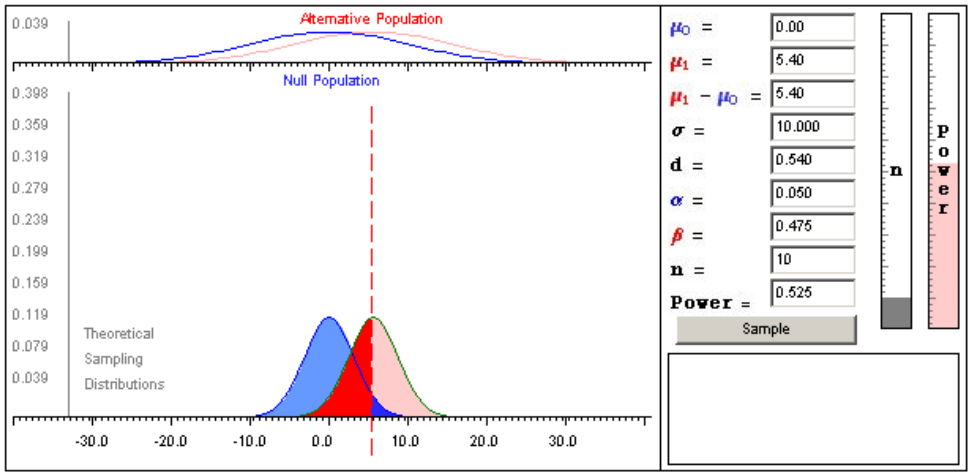
http://wise.cgu.edu/power/power_applet.html

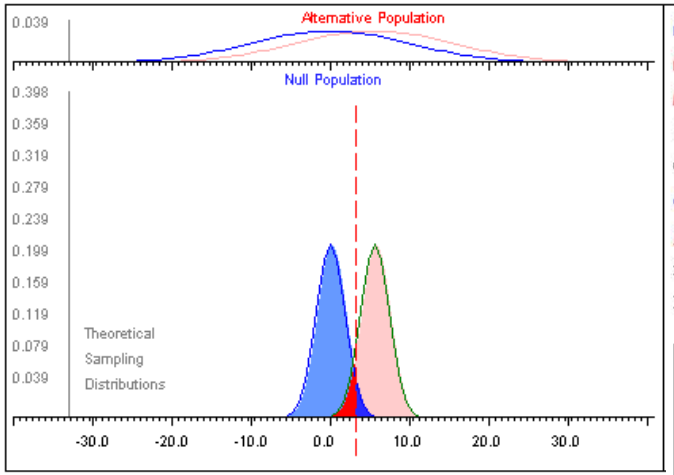
http://wise.cgu.edu/power/powerapplet1.html

WISE Power Applet

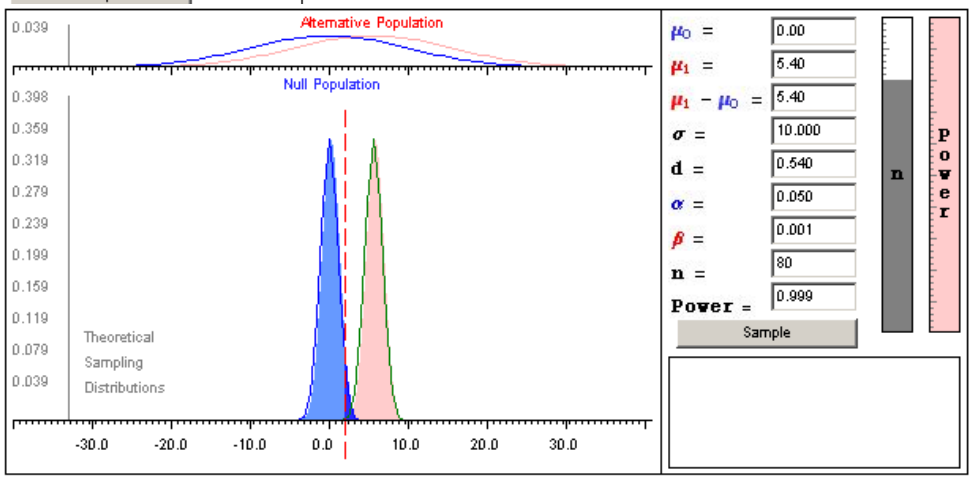
H_o: $\mu = 0$
$\sigma = 10$
Real $\mu = 5.4$
$\alpha = 5\%$

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



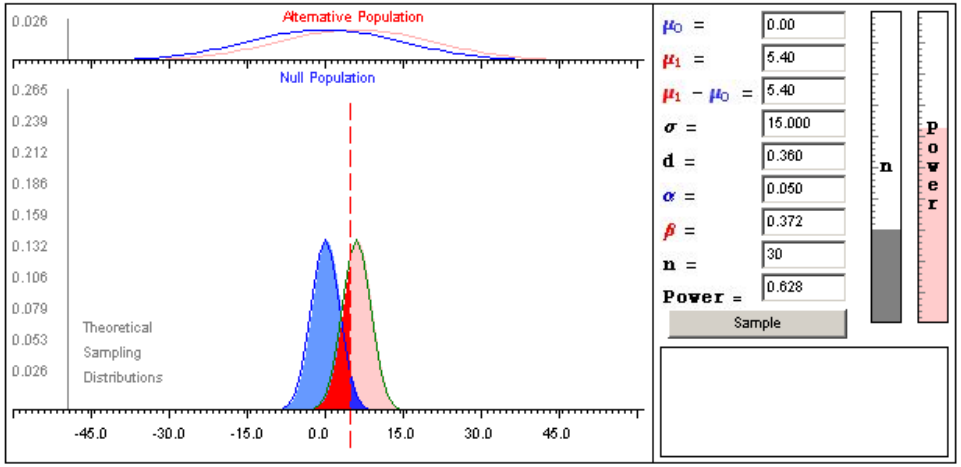


➔ **larger sample sizes yield greater power**



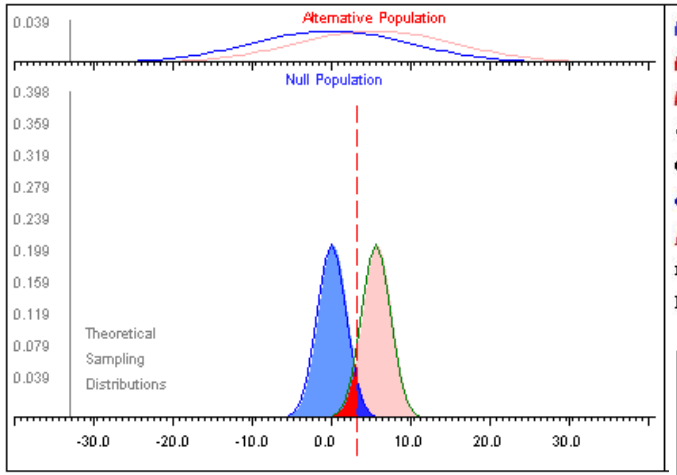
http://wise.cgu.edu/power/powerapplet1.html

WISE Power Applet

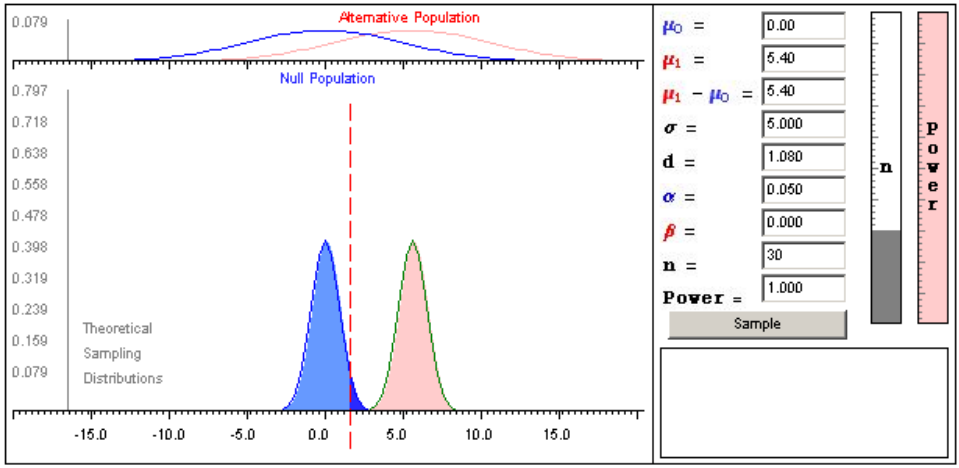


H_o: $\mu = 0$
Real $\mu = 5.4$
$n = 30$
$\alpha = 5\%$

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



➔ **smaller variability yields greater power**



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.

Section
Inference for the mean of a population

Change: Population s.d. sigma unknown.

- The ***t*** distribution
- One-sample ***t*** confidence interval
- One-sample ***t*** test
- **Matched pairs *t*** procedures
- Robustness of ***t*** procedures

Sweetening colas

Cola manufacturers want to test how much the sweetness of a new cola drink is affected by storage. The sweetness loss due to storage was evaluated by 10 professional tasters (by comparing the sweetness before and after storage):

Taster	Sweetness loss
1	2.0
2	0.4
3	0.7
4	2.0
5	−0.4
6	2.2
7	−1.3
8	1.2
9	1.1
10	2.3

Obviously, we want to test if storage results in a loss of sweetness, thus:

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

This looks familiar. However, here we do not know the population parameter σ.

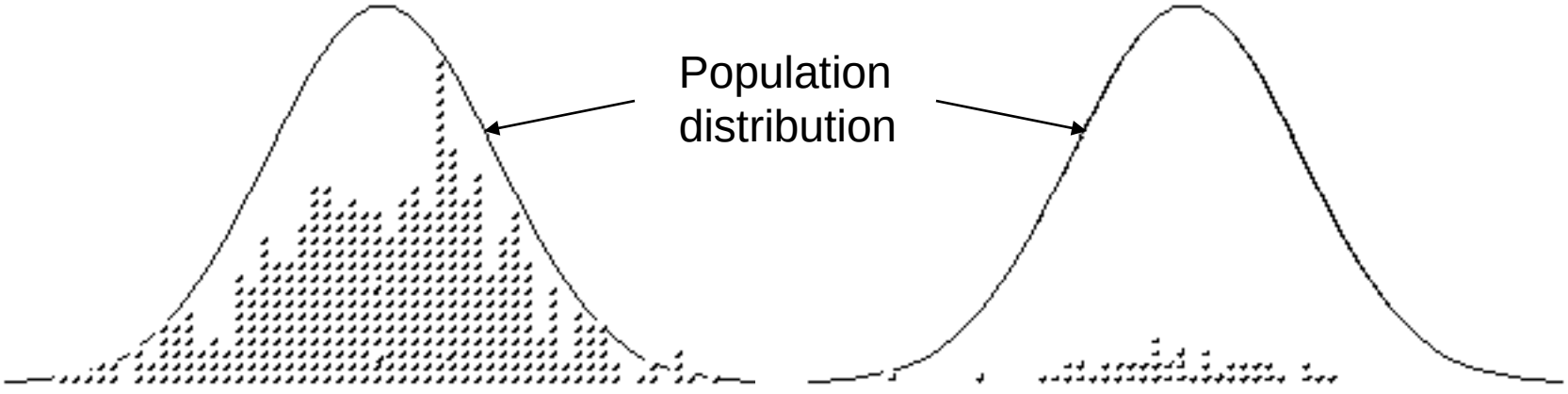
- The population of all cola drinkers is too large.
- Since this is a new cola recipe, we have no population data.

This situation is very common with real data.

When σ is unknown

The sample standard deviation *s* provides an estimate of the population standard deviation σ.

- When the sample size is large, the sample is likely to contain elements representative of the whole population. Then *s* is a good estimate of σ.
- But when the sample size is small, the sample contains only a few individuals. Then *s* is a more mediocre estimate of σ.



Standard deviation s – standard error $s/\sqrt{n}$

For a sample of size n,

the sample standard deviation s is:

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

$n - 1$ is the "degrees of freedom."

The value $s/\sqrt{n}$ is called the standard error of the sample mean or simply standard error of the mean (**SEM**).

Scientists often present sample results as mean ± SEM.

A study examined the effect of a new medication on the seated systolic blood pressure. The results, presented as mean ± SEM for 25 patients, are 113.5 ± 8.9.

What is the standard deviation s of the sample data?

$$\text{SEM} = s/\sqrt{n} \iff s = \text{SEM}*\sqrt{n}$$

$$s = 8.9*\sqrt{25} = 44.5$$

The *t* distribution:

The goal is to estimate or test for unknown ***μ*** in situation when ***σ*** is also unknown .

Solution: estimate ***σ*** by ***s*** and use intelligently in formulas.

Challenge: the distribution of the test statistic will change and will no longer be z-distribution.

Suppose that an SRS of size *n* is drawn from an $N(\mu, \sigma)$ population.

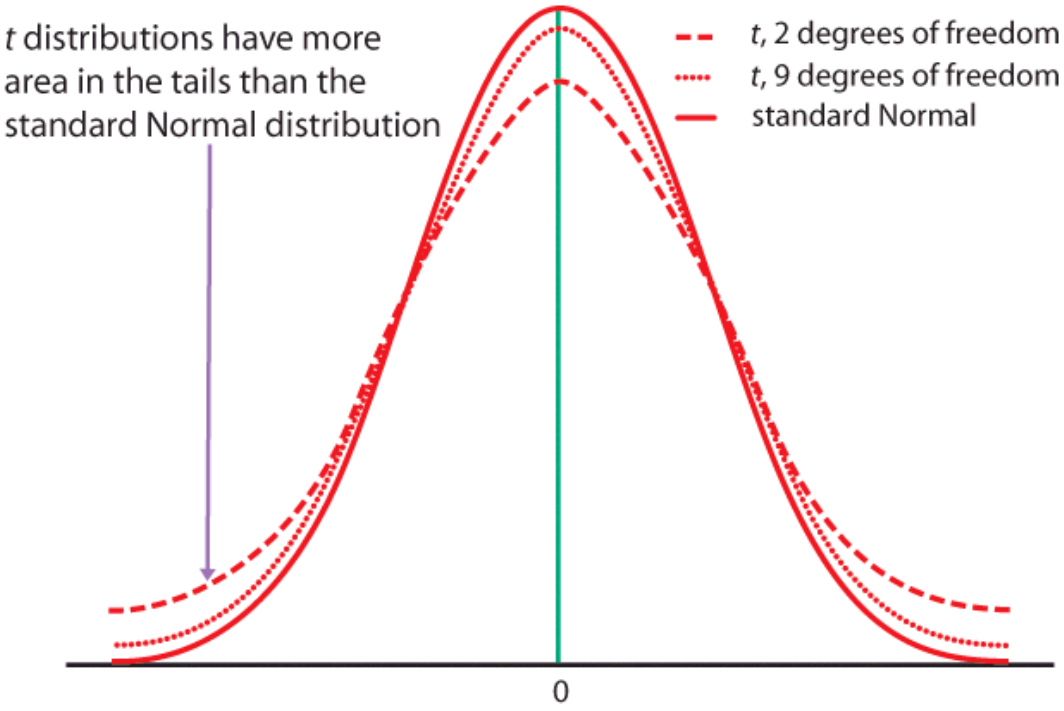
- When σ is known, the sampling distribution is $N(\mu, \sigma/\sqrt{n})$.

- When σ is estimated from the sample standard deviation *s*, the sampling distribution follows a ***t* distribution $t(\mu, s/\sqrt{n})$ with degrees of freedom $n - 1$.**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ is the **one-sample *t* statistic.**

When *n* is very large, *s* is a very good estimate of σ and the corresponding *t* distributions are very close to the normal distribution.

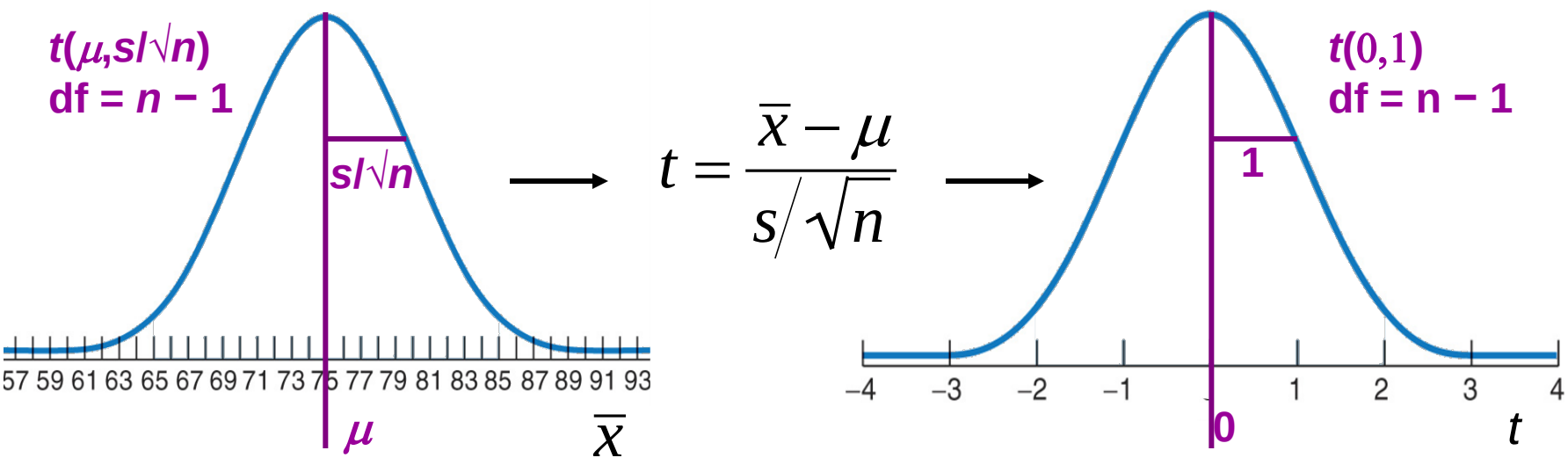
The *t* distributions become wider for smaller sample sizes, reflecting the lack of precision in estimating σ from *s*.



Standardizing the data

As with the normal distribution, the first step is to standardize the data.

Then we can use the **Table A.8** to obtain the area under the curve.



Here, μ is the mean (center) of the sampling distribution,
and the standard error of the mean s/√n is its standard deviation (width).
You obtain s, the standard deviation of the sample, with your calculator.

The one-sample *t*-confidence interval

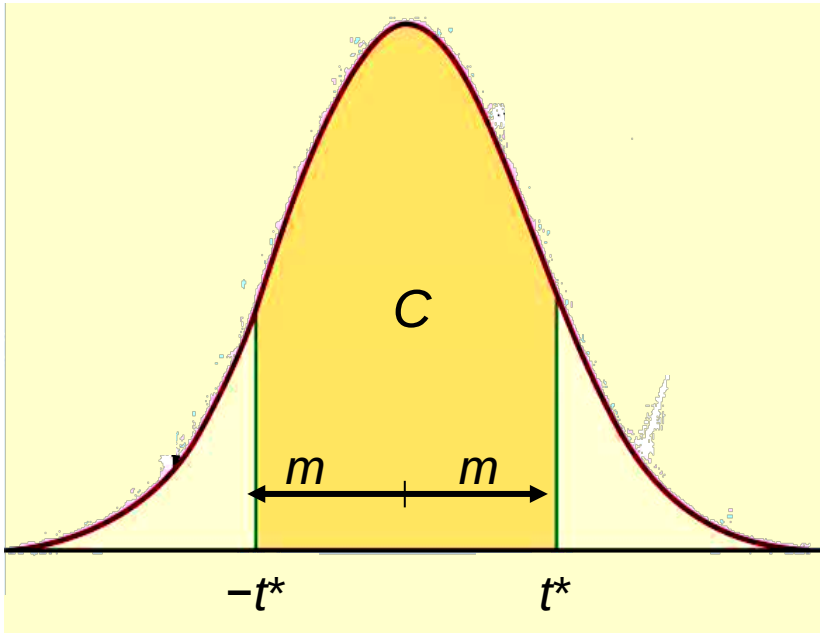
The **level *C* confidence interval** is an interval with probability *C* of containing the true population parameter.

We have a data set from a population with both μ and σ unknown. We use $\bar{x}$ to estimate μ, and *s* to estimate σ, using a *t* distribution (df n−1).

Practical use of *t* : *t

- *C* is the area between −*t** and *t**.
- We find *t** in the line of Table A.8 for nu = n−1 and probability (1-*C*)/2.
- The margin of error *m* is:

$$m = t^* s/\sqrt{n}$$

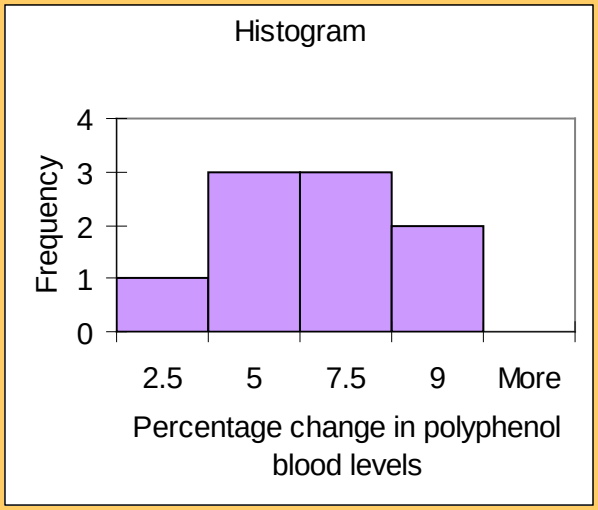


Red wine, in moderation

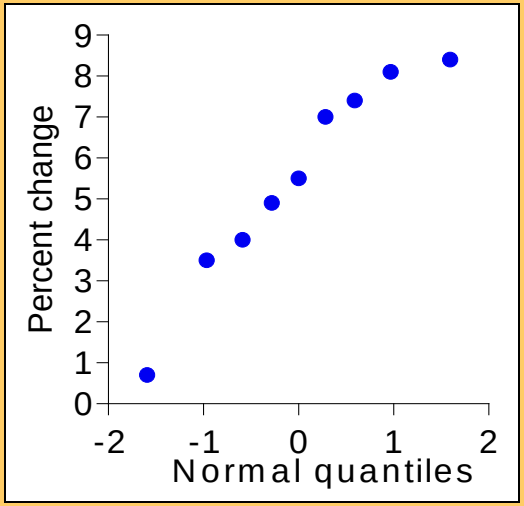
Drinking red wine in moderation may protect against heart attacks. The polyphenols it contains act on blood cholesterol and thus are a likely cause.

To see if moderate red wine consumption increases the average blood level of polyphenols, a group of nine randomly selected healthy men were assigned to drink half a bottle of red wine daily for two weeks. Their blood polyphenol levels were assessed before and after the study, and the percent change is presented here: **0.7 3.5 4 4.9 5.5 7 7.4 8.1 8.4**

Firstly: Are the data approximately normal?



0	7
1	
2	
3	5
4	09
5	5
6	
7	04
8	14



There is a low value, but overall the data can be considered reasonably normal.

What is the 95% confidence interval for the average percent change?

Sample average = 5.5; s = 2.517; df = $n - 1$ = 8

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
(…)												
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

The sampling distribution is a t distribution with $n - 1$ degrees of freedom.

For df = 8 and C = 95%, t^* = 2.306.

The margin of error m is: $m = t^*s/\sqrt{n} = 2.306*2.517/\sqrt{9} \approx 1.93$.

With 95% confidence, the population average percent increase in polyphenol blood levels of healthy men drinking half a bottle of red wine daily is between 3.6% and 7.6%. *Important: The confidence interval shows how large the increase is, but not if it can have an impact on men's health.*

The one-sample *t*-test

As in the previous chapter, a test of hypotheses requires a few steps:

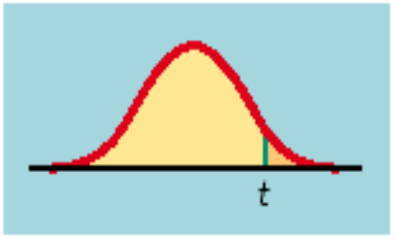
1. Stating the null and alternative hypotheses (H_0 versus H_a)
2. Deciding on a one-sided or two-sided test
3. Choosing a significance level α
4. Calculating t and its degrees of freedom
5. Finding the area under the curve *with Table A.8*
6. Stating the P-value and interpreting the result

The ***P-value*** is the probability, if H_0 is true, of randomly drawing a sample like the one obtained or more extreme, in the direction of H_a.

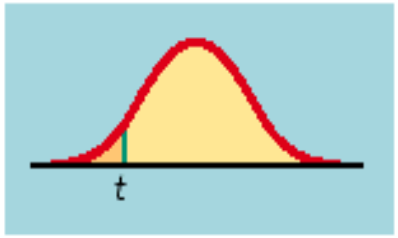
The P-value is calculated as the corresponding area under the curve, one-tailed or two-tailed depending on H_a:

One-sided (one-tailed)

$H_a: \mu > \mu_0 \Rightarrow P(T \geq t)$



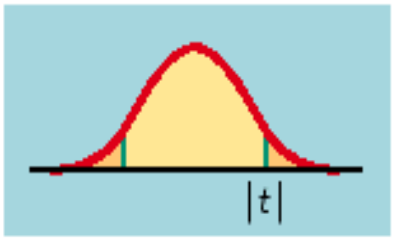
$H_a: \mu < \mu_0 \Rightarrow P(T \leq t)$



$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Two-sided (two-tailed)

$H_a: \mu \neq \mu_0 \Rightarrow 2P(T \geq |t|)$



- The table A.8 contains only a few values.
- To find exact p-values use R
- The function to be used is:
  - pt(quantile, df)

Sweetening colas (*continued*)

Is there evidence that storage results in sweetness loss for the new cola recipe at the 0.05 level of significance (α = 5%)?

H_0: $\mu = 0$ versus H_a: $\mu > 0$ (one-sided test)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.02 - 0}{1.196/\sqrt{10}} = 2.70$$

- The critical value t_α = 1.833.
  $t > t_\alpha$ thus the result is significant.
- 2.398 < t = 2.70 < 2.821 thus 0.02 > p > 0.01.
  $p < \alpha$ thus the result is significant.

Taster	Sweetness loss
1	2.0
2	0.4
3	0.7
4	2.0
5	-0.4
6	2.2
7	-1.3
8	1.2
9	1.1
10	2.3
Average	**1.02**
Standard deviation	**1.196**
Degrees of freedom	***n* − 1 = 9**

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
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

The t-test has a significant p-value. We reject H_0.

There is a significant loss of sweetness, on average, following storage.

Sweetening colas (*continued*)

Session

Minitab

One-Sample T: Loss

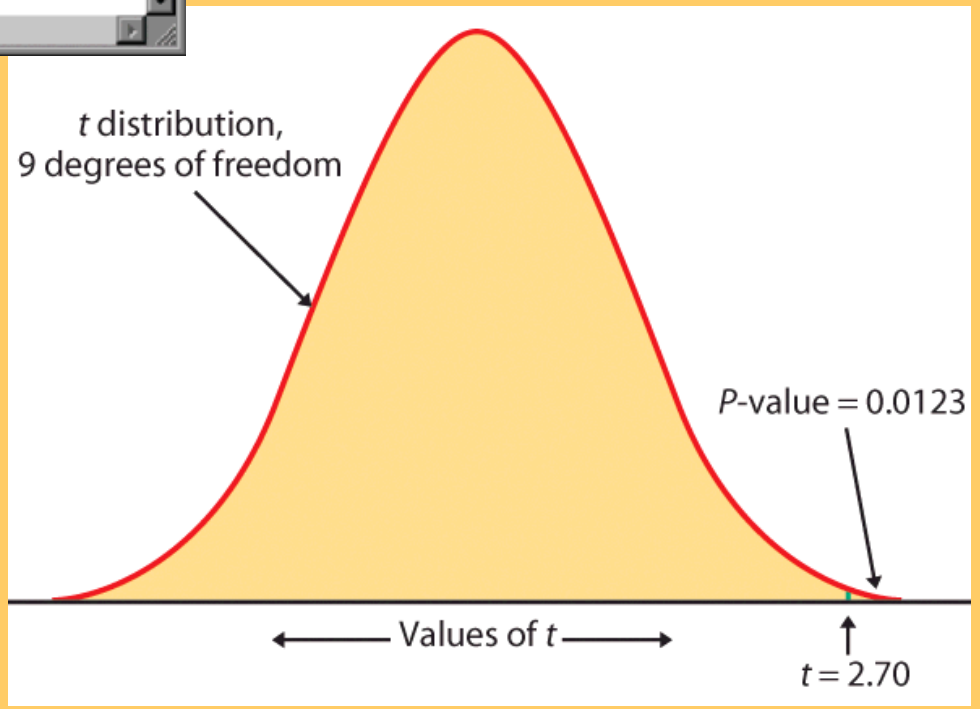
Test of mu = 0 vs mu > 0

Variable	N	Mean	StDev	SE Mean
Loss	10	1.020	1.196	0.378

Variable	95.0% Lower Bound	T	P
Loss	0.327	2.70	0.012

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.02 - 0}{1.196/\sqrt{10}} = 2.70$$

$$df = n - 1 = 9$$



In R, you can obtain the precise p-value once you have calculated *t*:

Using the function pt(2.7, 9)

which gives 0.9878032 and taking 1- this value (WHY?) we obtain 0.01219685

Matched pairs *t* procedures

Sometimes we want to compare treatments or conditions at the individual level. These situations produce two samples that are not independent — they are related to each other. The members of one sample are identical to, or matched (paired) with, the members of the other sample.

- Example: Pre-test and post-test studies look at data collected on the same sample elements before and after some experiment is performed.
- Example: Twin studies often try to sort out the influence of genetic factors by comparing a variable between sets of twins.
- Example: Using people matched for age, sex, and education in social studies allows canceling out the effect of these potential lurking variables.

In these cases, we use the paired data to test the difference in the two population means. The variable studied becomes $X_{difference} = (X_1 - X_2)$, and

$$H_0: \mu_{difference} = 0 \; ; \; H_a: \mu_{difference} > 0 \text{ (or } < 0, \text{ or } \neq 0)$$

Conceptually, this is not different from tests on one population.

Sweetening colas *(revisited)*

The sweetness loss due to storage was evaluated by 10 professional tasters (comparing the sweetness **before and after** storage):

Taster	Sweetness loss
1	2.0
2	0.4
3	0.7
4	2.0
5	−0.4
6	2.2
7	−1.3
8	1.2
9	1.1
10	2.3

We want to test if storage results in a loss of sweetness, thus:

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

Although the text didn't mention it explicitly, this is a pre-/post-test design and the variable is the difference in cola sweetness before minus after storage.

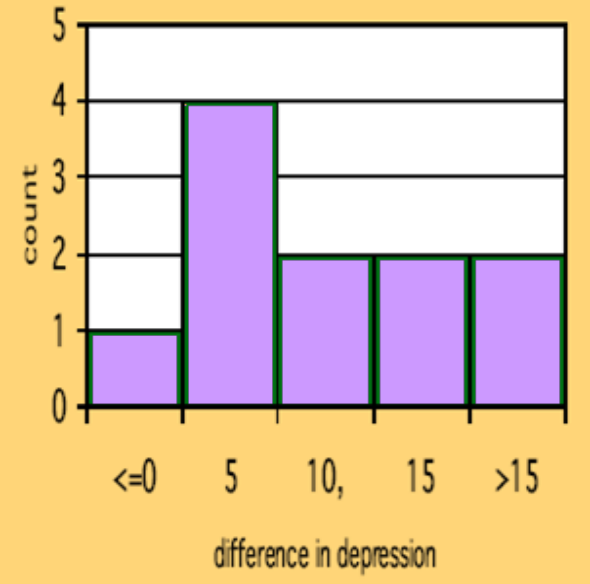
A matched pairs test of significance is indeed just like a one-sample test.

Does lack of caffeine increase depression?

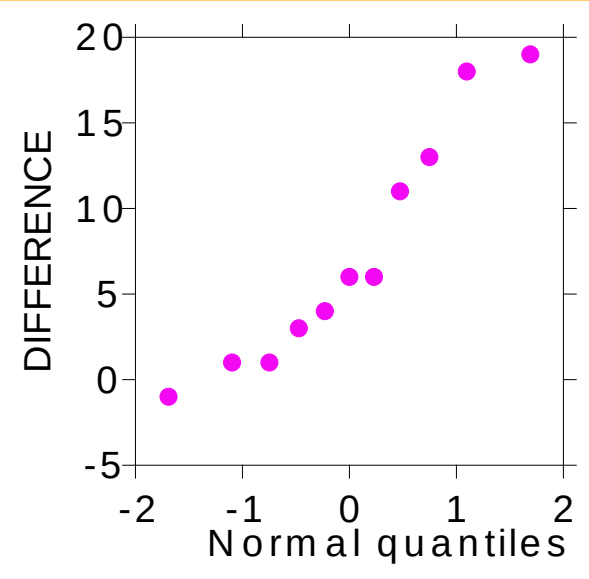
Individuals diagnosed as caffeine-dependent are deprived of caffeine-rich foods and assigned to receive daily pills. Sometimes, the pills contain caffeine and other times they contain a placebo. Depression was assessed.

Subject	Depression with Caffeine	Depression with Placebo	Placebo - Cafeine
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	-1

- There are 2 data points for each subject, but we'll only look at the difference.
- The sample distribution appears appropriate for a *t*-test.



11 "difference" data points.



Does lack of caffeine increase depression?

For each individual in the sample, we have calculated a difference in depression score (placebo minus caffeine).

There were 11 "difference" points, thus df = $n - 1 = 10$.

We calculate that $\bar{x} = 7.36$; $s = 6.92$

Subject	Depression with Caffeine	Depression with Placebo	Placebo - Cafeine
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	-1

H_0: $\mu_{difference} = 0$; H_0: $\mu_{difference} > 0$

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{7.36}{6.92/\sqrt{11}} = 3.53$$

For df = 10, $3.169 < t = 3.53 < 3.581$ ➔ $0.005 > p > 0.0025$

Caffeine deprivation causes a significant increase in depression.

Robustness

The *t* procedures are exactly correct when the population is distributed exactly normally. However, most real data are not exactly normal.

The *t* procedures are **robust** to small deviations from normality – the results will not be affected too much. Factors that strongly matter:

- **Random sampling.** The sample **must** be an SRS from the population.
- **Outliers and skewness.** They strongly influence the mean and therefore the *t* procedures. However, their impact diminishes as the sample size gets larger because of the Central Limit Theorem.

Specifically:

- When $n < 15$, the data must be close to normal and without outliers.
- When $15 > n > 40$, mild skewness is acceptable but not outliers.
- When $n > 40$, the *t*-statistic will be valid even with strong skewness.

Power of the *t*-test

The power of the one sample *t*-test against a specific alternative value of the population mean μ assuming a fixed significance level α is the probability that the test will reject the null hypothesis when the alternative is true.

Calculation of the exact power of the *t*-test is a bit complex. But an approximate calculation that acts as if σ were known is almost always adequate for planning a study. This calculation is very much like that for the *z*-test.

> When guessing σ, it is always better to err on the side of a standard deviation that is a little larger rather than smaller. We want to avoid a failing to find an effect because we did not have enough data.

Does lack of caffeine increase depression?

Suppose that we wanted to perform a similar study but using subjects who regularly drink caffeinated tea instead of coffee. For each individual in the sample, we will calculate a difference in depression score (placebo minus caffeine). How many patients should we include in our new study?

In the previous study, we found that the average difference in depression level was 7.36 and the standard deviation 6.92.

- We will use $\mu = 3.0$ as the alternative of interest. We are confident that the effect was larger than this in our previous study, and this amount of an increase in depression would still be considered important.
- We will use $s = 7.0$ for our guessed standard deviation.

We can choose a one-sided alternative because, like in the previous study, we would expect caffeine deprivation to have negative psychological effects.

Does lack of caffeine increase depression?

How many subjects should we include in our new study? Would 16 subjects be enough? Let's compute the power of the *t*-test for

$$H_0: \mu_{\text{difference}} = 0 \; ; \; H_a: \mu_{\text{difference}} > 0$$

against the alternative $\mu = 3$. For a significance level α 5%, the *t*-test with *n* observations rejects H_0 if *t* exceeds the upper 5% significance point of *t*(df:15) = 1.729. For $n = 16$ and $s = 7$:

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{\bar{x}}{7/\sqrt{16}} \geq 1.753 \;\Rightarrow\; \bar{x} \geq 1.06775$$

The power for $n = 16$ would be the probability that $\bar{x} \geq 1.068$ when $\mu = 3$, using $\sigma = 7$. Since we have σ, we can use the normal distribution here:

$$P(\bar{x} \geq 1.068 \text{ when } \mu = 3) = P\left(z \geq \frac{1.068 - 3}{7/\sqrt{16}}\right)$$

$$= P(z \geq -1.10) = 1 - P(z \leq -1.10) = 0.8643$$

The power would be about 86%.

Inference for non-normal distributions

What if the population is clearly non-normal and your sample is small?

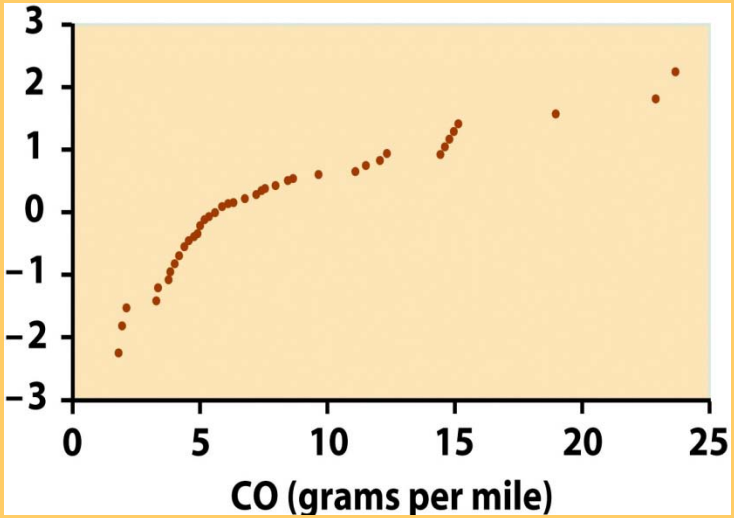
- If the data are skewed, you can attempt to **transform** the variable to bring it closer to normality (e.g., logarithm transformation). The *t*-procedures applied to transformed data are quite accurate for even moderate sample sizes.
- A distribution other than a normal distribution might describe your data well. Many non-normal models have been developed to provide inference procedures too.
- You can always use a **distribution-free ("nonparametric")** inference procedure (more in a second class in statistics) that does not assume any specific distribution for the population. But it is usually less powerful than distribution-driven tests (e.g., *t* test).

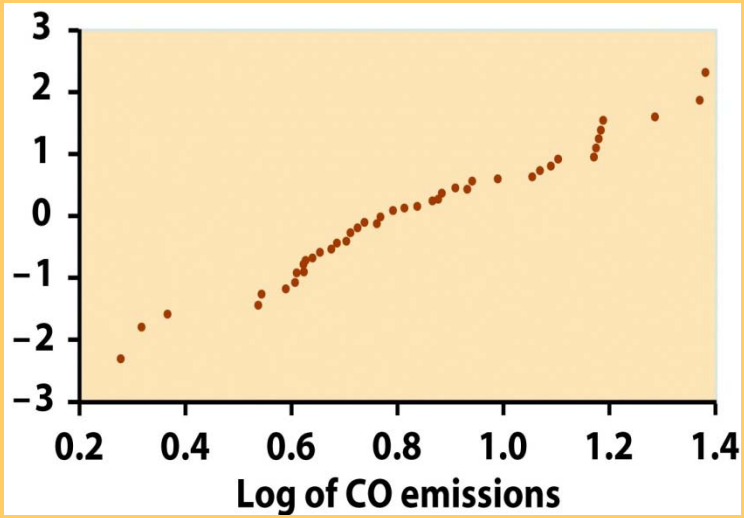
Transforming data

The most common transformation is the **logarithm (log),** which tends to pull in the right tail of a distribution.

Instead of analyzing the original variable *X*, we first compute the logarithms and analyze the values of log *X*.

However, we cannot simply use the confidence interval for the mean of the logs to deduce a confidence interval for the mean μ in the original scale.





Normal quantile plots for 46 car CO emissions

Nonparametric method: the sign test

A distribution-free test usually makes a statement of hypotheses about the median rather than the mean (e.g., “are the medians different”). This makes sense when the distribution may be skewed.

H_0: population median = 0 vs. H_a: population median > 0

A simple distribution-free test is the **sign test for matched pairs.**

Calculate the matched difference for each individual in the sample.
Ignore pairs with difference 0.
The number of trials n is the count of the remaining pairs.
The test statistic is the count X of pairs with a positive difference.
P-values for X are based on the binomial $B(n, 1/2)$ distribution.

H_0: $p = 1/2$ vs. H_a: $p > 1/2$

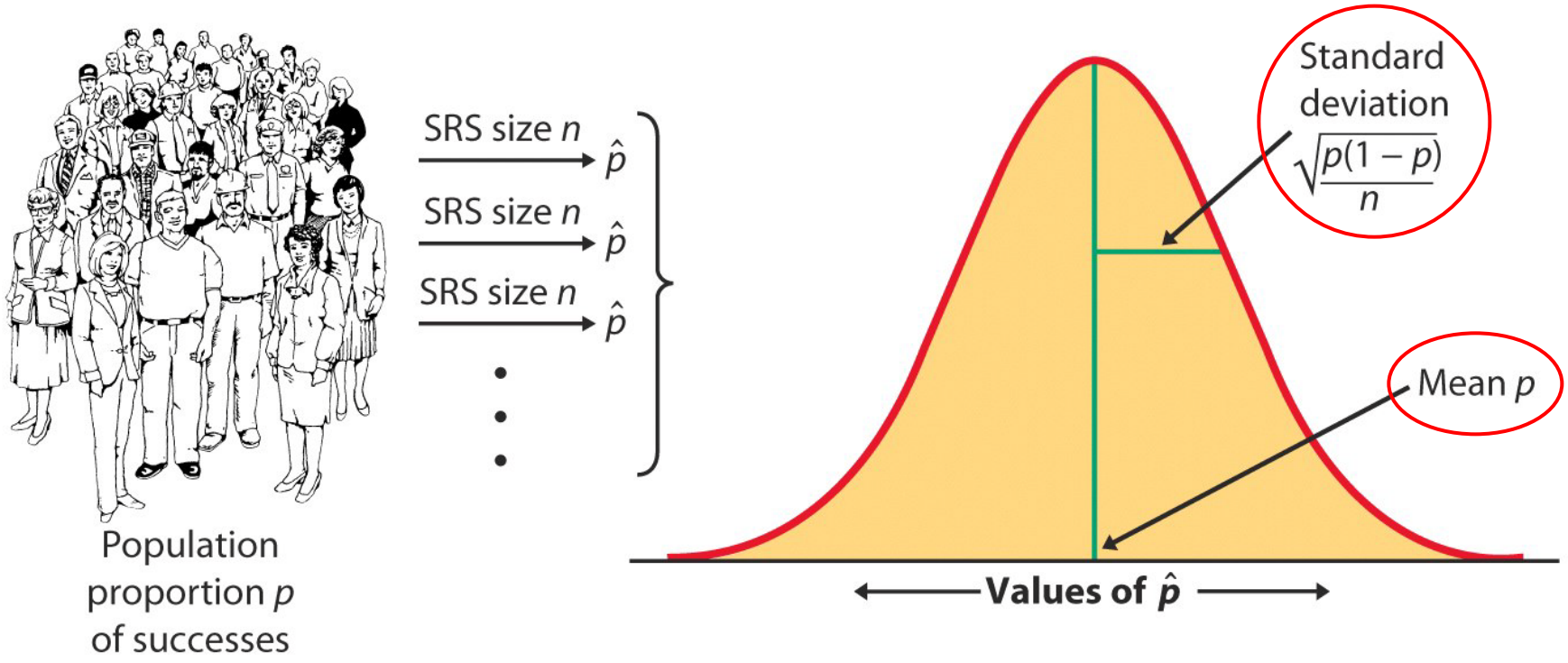
Section

- Inference for a Single Proportion:
- ✓ confidence intervals, planning sample size for a given margin of error
- ✓ test of significance for a single proportion

This will be very similar to what we did for means previously

Sampling distribution of $\hat{p}$ — reminder

The sampling distribution of a sample proportion $\hat{p}$ is approximately normal (normal approximation of a binomial distribution) when the sample size is large enough.



Conditions for inference on p

Assumptions:

1. The data used for the estimate are an SRS from the population studied.

2. The population is at least 10 times as large as the sample used for inference. This ensures that the standard deviation of $\hat{p}$ is close to $\sqrt{p(1-p)/n}$

3. The sample size *n* is large enough that the sampling distribution can be approximated with a normal distribution. How large a sample size is required depends in part on the value of *p* and the test conducted. Otherwise, rely on the binomial distribution.

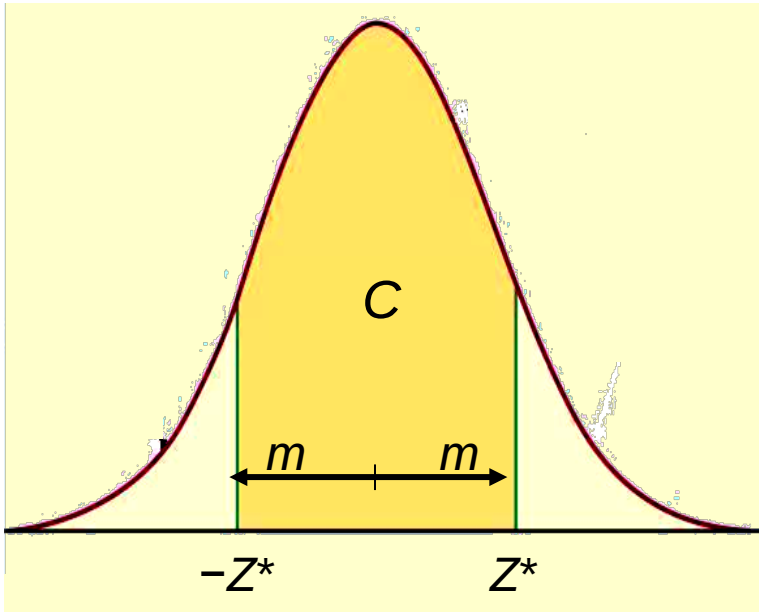
Large-sample confidence interval for *p*

Confidence intervals contain the population proportion *p* in *C*% of samples. For an SRS of size *n* drawn from a large population and with sample proportion $\hat{p}$ calculated from the data, an **approximate level *C* confidence interval** for *p* is:

$$\hat{p} \pm m, \quad m \text{ is the margin of error}$$

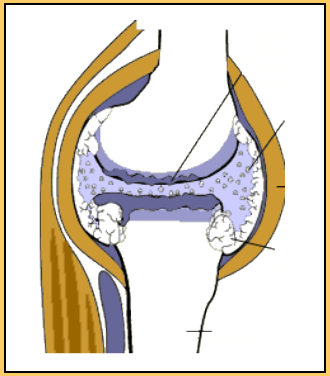
$$m = z^* SE = z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

Use this method when the number of successes and the number of failures are both at least 15.



C is the area under the standard normal curve between −*z** and *z**.

Medication side effects



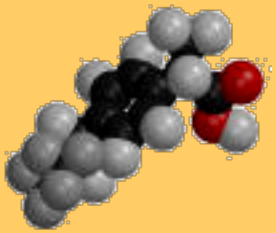
Arthritis is a painful, chronic inflammation of the joints. An experiment on the side effects of pain relievers examined arthritis patients to find the proportion of patients who suffer side effects.

What are some side effects of ibuprofen?

Serious side effects (seek medical attention immediately):

- Allergic reaction (difficulty breathing, swelling, or hives),
- Muscle cramps, numbness, or tingling,
- Ulcers (open sores) in the mouth,
- Rapid weight gain (fluid retention),
- Seizures,
- Black, bloody, or tarry stools,
- Blood in your urine or vomit,
- Decreased hearing or ringing in the ears,
- Jaundice (yellowing of the skin or eyes), or
- Abdominal cramping, indigestion, or heartburn,

Less serious side effects (discuss with your doctor):



- Dizziness or headache,
- Nausea, gaseousness, diarrhea, or constipation,
- Depression,
- Fatigue or weakness,
- Dry mouth, or
- Irregular menstrual periods

Let's calculate a 90% confidence interval for the population proportion of arthritis patients who suffer some "adverse symptoms."

What is the sample proportion $\hat{p}$?

$$\hat{p} = \frac{23}{440} \approx 0.052$$

What is the sampling distribution for the proportion of arthritis patients with adverse symptoms for samples of 440?

$$\hat{p} \approx N\left(p, \sqrt{p(1-p)/n}\right)$$

For a 90% confidence level, $z^* = 1.645$.

z*	0.67	0.841	1.036	1.282	1.645	1.960	2.054	2.326
	50%	60%	70%	80%	90%	95%	96%	98%
					Confidence level C			

Using the large sample method, we calculate a margin of error m:

$$m = z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

$$m = 1.645 * \sqrt{0.052(1-0.052)/440}$$

$$m = 1.645 * 0.014 \approx 0.023$$

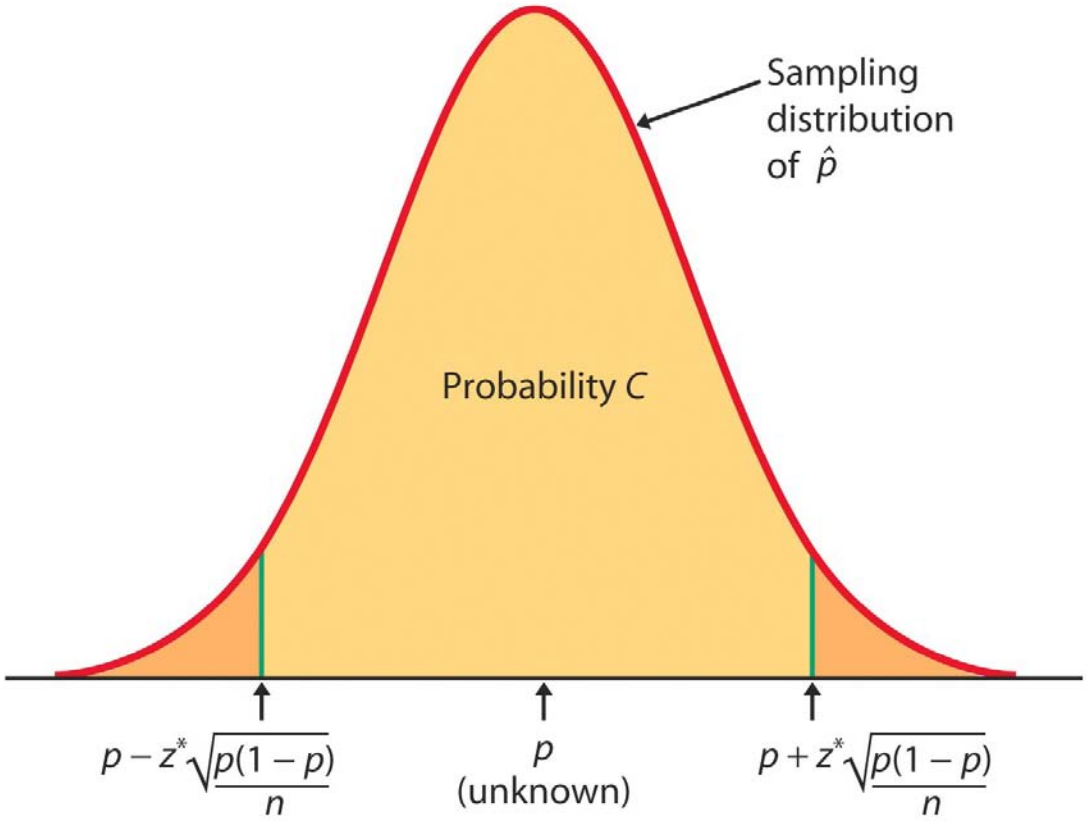
$$90\%\ \text{CI for } p: \hat{p} \pm m$$

$$\text{or } 0.052 \pm 0.023$$

➔ With 90% confidence level, between 2.9% and 7.5% of arthritis patients taking this pain medication experience some adverse symptoms.

Because we have to use an estimate of *p* to compute the margin of error, confidence intervals for a population proportion are not very accurate.

$$m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



Specifically, we tend to be incorrect more often than the confidence level would indicate. But there is no systematic amount (because it depends on p).

Use with caution!

Interpretation: magnitude vs. reliability of effects

The **reliability** of an interpretation is related to the strength of the evidence. The smaller the **p-value,** the stronger the evidence against the null hypothesis and the more confident you can be about your interpretation.

The **magnitude** or **size** of an effect relates to the real-life relevance of the phenomenon uncovered. The p-value does NOT assess the relevance of the effect, nor its magnitude.

A **confidence interval** will assess the magnitude of the effect. However, magnitude is not necessarily equivalent to how theoretically or practically relevant an effect is.

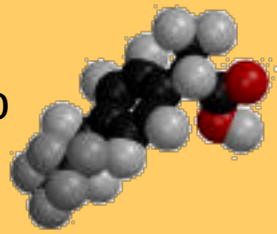
Sample size for a desired margin of error

You may need to choose a sample size large enough to achieve a specified margin of error. However, because the sampling distribution of $\hat{p}$ is a function of the population proportion *p,* this process requires that you guess a likely value for *p*: *p**.

$$p \sim N\left(p, \sqrt{p(1-p)/n}\right) \Rightarrow \boxed{m = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)}$$

The margin of error will be less than or equal to *m* if *p** is chosen to be 0.5.

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples.



What sample size would we need in order to achieve a margin of error no more than 0.01 (1%) for a 90% confidence interval for the population proportion of arthritis patients who suffer some “adverse symptoms.”

We could use 0.5 for our guessed p^*. However, since the drug has been approved for sale over the counter, we can safely assume that no more than 10% of patients should suffer “adverse symptoms” (a better guess than 50%).

For a 90% confidence level, $z^* = 1.645$.

z*	0.67	0.841	1.036	1.282	1.645	1.960	2.054	2.326
	50%	60%	70%	80%	90%	95%	96%	98%
					Confidence level C			

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.645}{0.01}\right)^2 (0.1)(0.9) \approx 2434.4$$

➔ To obtain a margin of error no more than 1%, we would need a sample size n of at least 2435 arthritis patients.