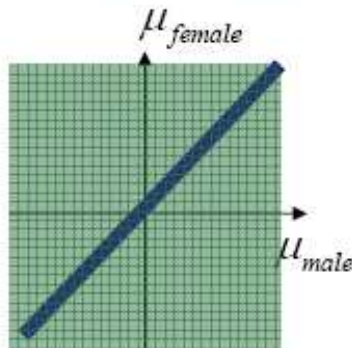


Hypothesis are ALWAYS about parameter regions

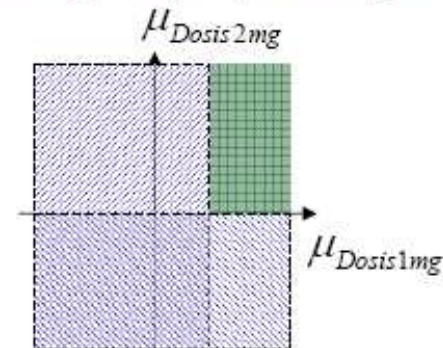
$$H_0 : \mu_{male} = \mu_{female}$$

$$H_1 : \mu_{male} \neq \mu_{female}$$



$$H_0 : \{ \mu_{Dosis1mg} \leq 2 \} \cup \{ \mu_{Dosis2mg} < 0 \}$$

$$H_1 : \{ \mu_{Dosis1mg} > 2 \} \cap \{ \mu_{Dosis2mg} \geq 0 \}$$



Hypothesis are NEVER about specific realizations of the random variable

Joe is a hypertense patient. Research hypothesis: Given our previous study, the effect of 1mg of the drug will have a positive effect larger than 2 on Joe.

Joe and Mary have been married for then 5 years. Research hypothesis: Joe will report higher intimacy level than Mary.

Hypothesis NEVER use “Not all”, “Some”, “None”, “All”

Research hypothesis: All hypertense patients benefit from a new drug.

Research hypothesis: None hypertense patients benefit from a new drug.

Problem: We would have to measure absolutely ALL hypertense patients

Research hypothesis: Not all hypertense patients benefit from a new drug.

Research hypothesis: Some hypertense patients benefit from a new drug.

Problem: Too imprecise, being true does not provide much information

What can we do with hypothesis testing?

- You **CAN** reject the null hypothesis and accept the alternative hypothesis
- You **CAN** fail to reject the null hypothesis because, there is not sufficient evidence to reject it
- You **CANNOT** accept the null hypothesis and reject the alternative because you would need to measure absolutely all elements (for instance, all couples married for more than 5 years).

It's like in legal trials:

- The null hypothesis is the innocence of the defendant.
- You **CAN** reject his innocence based on proofs (always with a certain risk).
- You **CAN** fail to reject his innocence.
- You **CANNOT** prove his innocence (you would need absolutely all facts)

The goal of hypothesis testing is to disprove the null hypothesis! We do this by proving that if the null hypothesis were true, then there would be a very low probability of observing the sample we have actually observed.

However, there is always the risk that we have been unlucky with our sample, this is our confidence level (the p-value is also related to this risk: the lower the p-value, the lower the risk).

4.3 Assumptions about hypothesis tests

- Ideally the violation of the test assumptions invalidate the test result.
- However, there are tests that are more robust than others to violations.
- And the same test may be more robust to violations in one assumption than another.

That's why it is better to know the assumptions about:

- The population distribution
- Residuals
- The nature of the sample
- Measurement errors
- The model being tested.

Assumptions about the population

- Parametric tests:
 - usually make an assumption about the population (normally, normality).
 - Tests: Shapiro-Wilks (1D), Kolmogorov-Smirnov(1D), Smith-Jain (nD)
 - Samples are independent and identically distributed
 - Homoscedasticity is also usually assumed :
 - Regression: the variance of the dependent variables is the same across the range of predictor variables
 - ANOVA-like: the variance of the variable studied is the same among groups
 - Tests: Levene, Breusch-Pagan, White
 - From these assumptions the statistic distribution is derived.
 - Normally applied to ratio/interval variables.

Assumptions about the population

- Nonparameteric tests:
 - Don't make any "parametric assumption"
 - Samples are independent and identically distributed
 - Normally applied to ordinal/categorical variables
 - Permutation tests assume that labels can be permuted

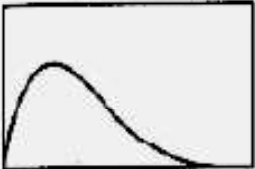


Normality

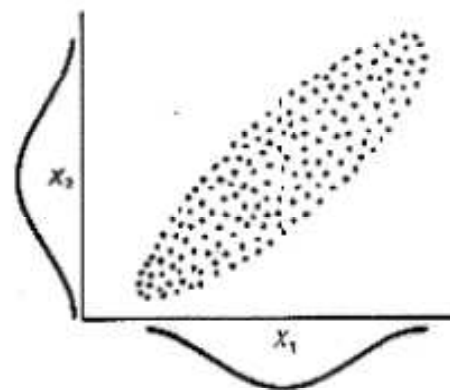
The solution to most assumption violations is provided by data transformations.

Homoscedasticity

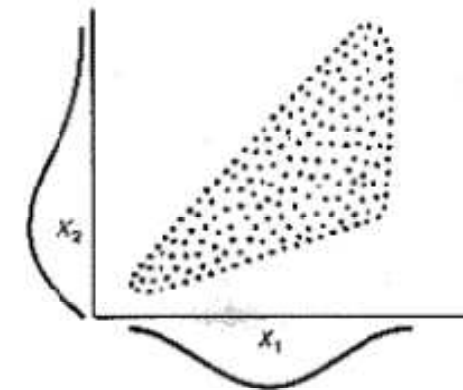
Sometimes can be detected by bare eye.

Table of sample pdfs and suggested transformation

Form	Transformation
	Square Root $Y = \sqrt{X}$
	Logarithm $Y = \log X$
	Inverse $Y = \frac{1}{X}$

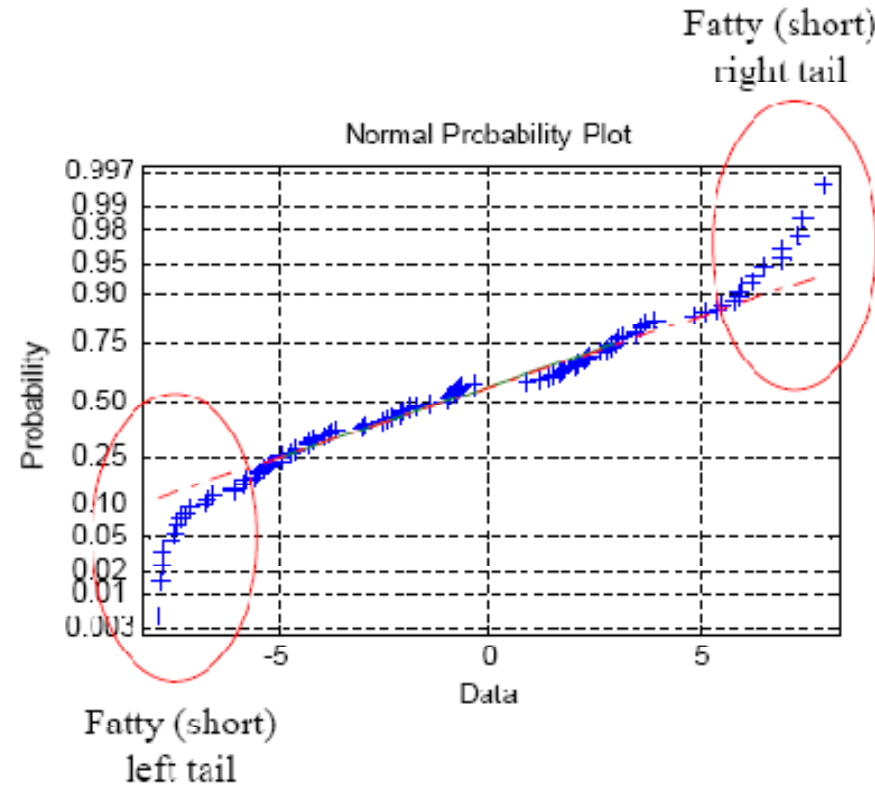
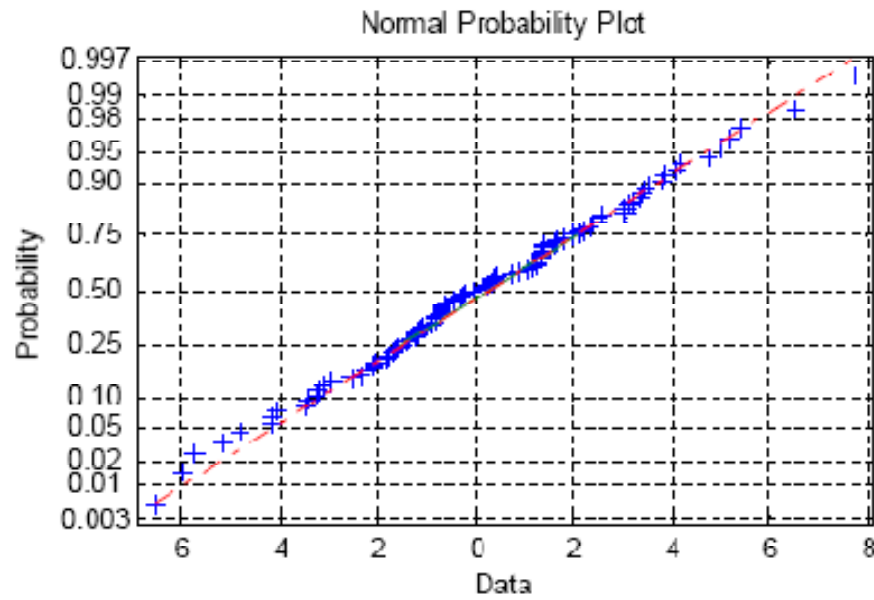


Homoscedasticity with both variables normally distributed



Heteroscedasticity with skewness on one variable

Normality: Probability plots



Assumptions about residuals (see also Chapter 8)

$$\text{SystolicPressure} = 163.7 + \alpha_{\text{treatment}} + \varepsilon$$

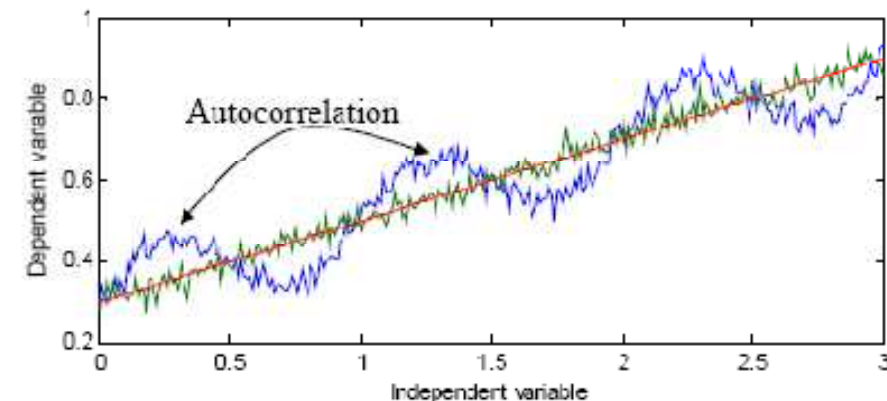
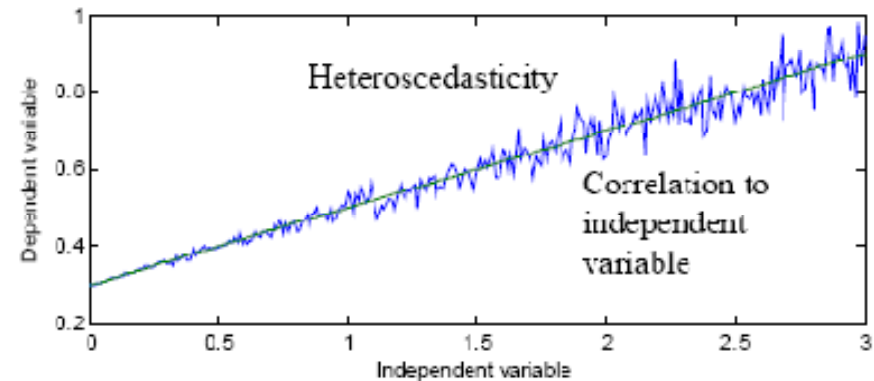
$$\text{Score} = \mu_{\text{RealMadrid}} + \alpha_{\text{strategy}} + \alpha_{\text{Raul}} + \alpha_{\text{Raul, strategy}} + \varepsilon$$

➤ Residuals

- Zero mean
- Homoscedasticity
- No autocorrelation
- Not correlated to the independent variable

➤ Solution:

- Change your model
- Use non parametric test

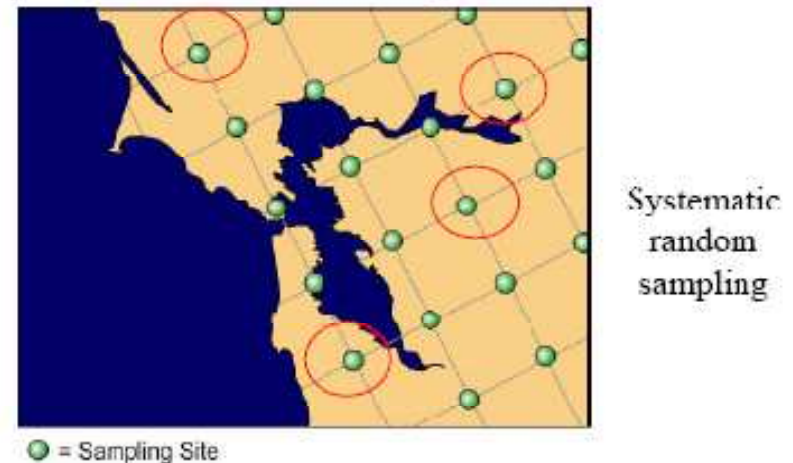
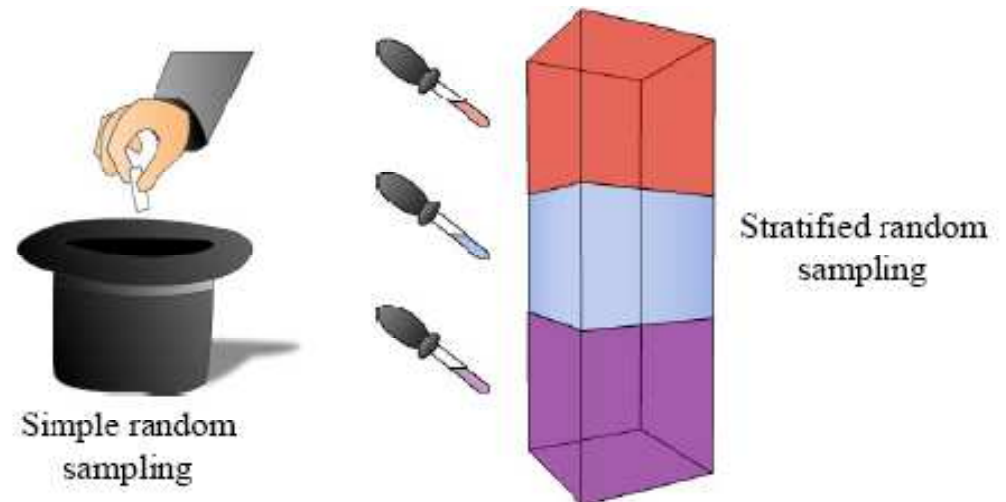


Assumptions about the nature of the sample

- Sample
 - Simple random sampling
 - Samples are independent

- Solution:
 - Random sampling but not simple → readjust variances
 - Sampling not random → inference cannot be used!
You cannot generalize to the whole population

Sampling: A study carried on volunteers
Violates random sampling



Assumptions about variables, measurements and models

➤ Nature of variables

- Categorical: Sex
- Ordinal: Degree of preference
- Interval: Temperature
- Ratio: Length

➤ Measurements

- Regression assumes that the independent variable (X) is measured without error, but the dependent variable is measured with error. If there is error in X, then association measures are underestimated.
- Check if your measurements are reliable.

➤ Models

- Many models assume linearity (for instance, the correlation coefficient)
- We assume that our model fully explain the phenomenon

Purchase Warm Clothes ↓ —————→ *Purchase Ice Creams* ↑

4.4 Examples

Coca-cola example

An engineer works for Coca Cola. He knows that the filling machine has a variance of 6 cl. (the filling process can be approximated by a Gaussian). Knowing this, he sets the machine to a target fill of 348 cl ($=330+3*6$). In a routine check with 25 cans, he measures an average of 345 cl. Is it possible that the machine is malfunctioning?



Step 1. Define your hypothesis

$$H_0 : \mu = 348$$

$$H_1 : \mu \neq 348$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{25}}} \sim N(0,1)$$

Step 3. Plug-in observed data

$$z = \frac{345 - 348}{\frac{6}{\sqrt{25}}} = -2.5$$

Step 4. Compute probabilities

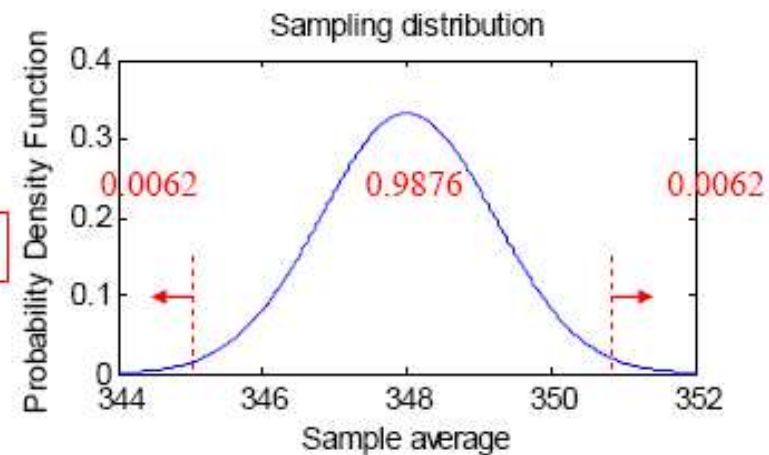
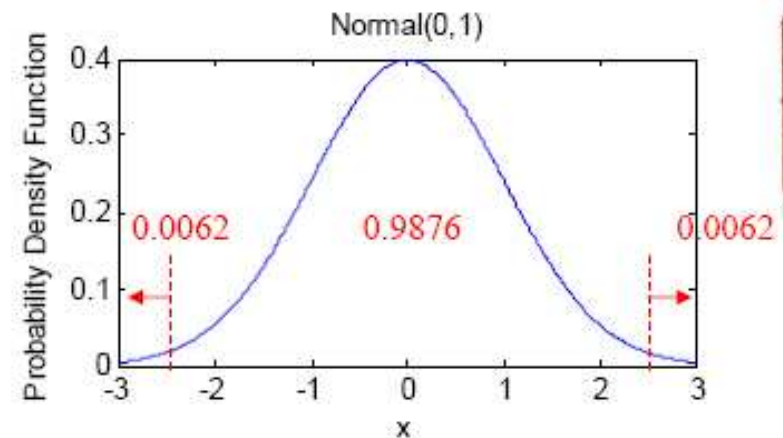
The probability of observing a sample average AT LEAST AS EXTREME AS THE OBSERVED is

$$P\{Z < -2.5\} + P\{Z > 2.5\} = 0.0062 \cdot 2 = 0.0124$$

Step 5. Reject or not H_0

Extremes happen: normally in 1.24% of the cases I will observe a sample average that is at least as far from 348 as 345.

p -value



$0.0124 < 0.05$ I will take the risk of not going for machine maintenance when I should have gone in 5% of the cases

$0.0124 > 0.01$ I will take the risk of not going for machine maintenance when I should have gone in 1% of the cases

More coca-cola

Is it possible that the machine is filling less than programmed?

Step 1. Define your hypothesis

$$H_0 : \mu \geq 348$$

$$H_1 : \mu < 348$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{25}}} \sim N(0,1)$$

Step 3. Plug-in observed data

$$z = \frac{345 - 348}{\frac{6}{\sqrt{25}}} = -2.5$$

Step 4. Compute probabilities

The probability of observing a sample average AT LEAST AS EXTREME AS THE OBSERVED is

$$P\{Z < -2.5\} = 0.0062$$

Step 5. Reject or not H_0

Extremes happen: normally in 1.24% of the cases I will observe a sample average that is at least as far from 348 as 345.

$$0.0062 < 0.01$$

I will take the risk of going for machine maintenance when I should have not in 1% of the cases

Ethnic example

A country that has 4 ethnics (Balzacs 40%, Crosacs 25%, Murads 30%, Isads 5%) imposes by law that medical schools accept students proportionally to the ethnics. The distribution of 1000 students admitted this year in a medical school has been 300, 220, 400 and 80. Is the school respecting the law?

Step 1. Define your hypothesis

$$H_0 : O_{Balzacs} = E_{Balzacs} \cap O_{Crosacs} = E_{Crosacs} \cap O_{Murads} = E_{Murads} \cap O_{Isads} = E_{Isads}$$

$$H_1 : O_{Balzacs} \neq E_{Balzacs} \cup O_{Crosacs} \neq E_{Crosacs} \cup O_{Murads} \neq E_{Murads} \cup O_{Isads} \neq E_{Isads}$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad X = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-1}^2$$

Step 3. Plug-in observed data

$$\begin{aligned} x &= \frac{(300 - 1000 \cdot 40\%)^2}{1000 \cdot 40\%} + \frac{(220 - 1000 \cdot 25\%)^2}{1000 \cdot 25\%} + \frac{(400 - 1000 \cdot 30\%)^2}{1000 \cdot 30\%} + \frac{(80 - 1000 \cdot 5\%)^2}{1000 \cdot 5\%} = \\ &= 25 + 3.6 + 33.3 + 18 = 79.9 \end{aligned}$$

Step 4. Compute acceptance region

$$P\{X > 79.9\} \approx 0$$

Step 5. Reject or not H_0

Reject!

Politics example

We want to know if there is any relationship between political affiliation and personality. We study 200 individuals obtaining the following data

	Democrat	Republican	Sum
Introvert	20	80	110 (55%)
Extrovert	50	40	90 (45%)
Sum	70 (35%)	120 (65%)	

Step 1. Define your hypothesis

$$H_0 : O_{Introvert, Democrat} = E_{Introvert, Democrat} \cap \dots \cap O_{Extrovert, Republican} = E_{Extrovert, Republican}$$

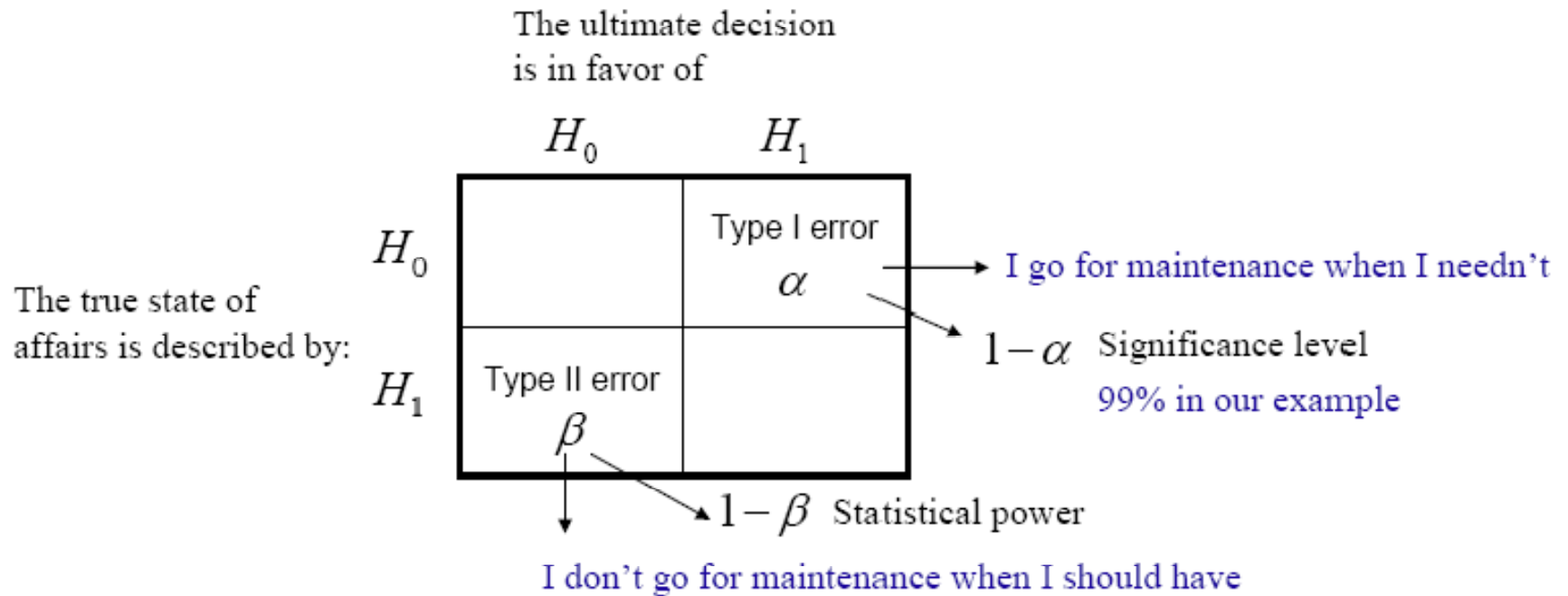
$$H_1 : O_{Introvert, Democrat} \neq E_{Introvert, Democrat} \cup \dots \cup O_{Extrovert, Republican} \neq E_{Extrovert, Republican}$$

Step 3. Plug-in observed data

$$\begin{aligned} x &= \frac{(30 - 200 \cdot 35\% \cdot 55\%)^2}{200 \cdot 35\% \cdot 55\%} + \frac{(80 - 200 \cdot 65\% \cdot 55\%)^2}{200 \cdot 65\% \cdot 55\%} + \frac{(50 - 200 \cdot 35\% \cdot 45\%)^2}{200 \cdot 35\% \cdot 45\%} + \frac{(40 - 200 \cdot 65\% \cdot 45\%)^2}{200 \cdot 65\% \cdot 45\%} \\ &= 1.88 + 1.01 + 10.87 + 5.85 = 19.61 \end{aligned}$$

Step 4. Compute acceptance region $P\{X > 19.61\} = 0.0002$

Decision making



Decision Rule: If $p - value < \alpha$, then reject H_0
 Otherwise, you cannot reject H_0

Drawback: it is exclusively driven by Type I errors

4.5 Use and abuse of tests

Causes 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 $0.4^{\circ}\text{Celsius}$ ($P\text{-value} < 0.01$). 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 is 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%:

$$H_0: \mu = 0 \text{ versus } H_a: \mu > 0$$

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?

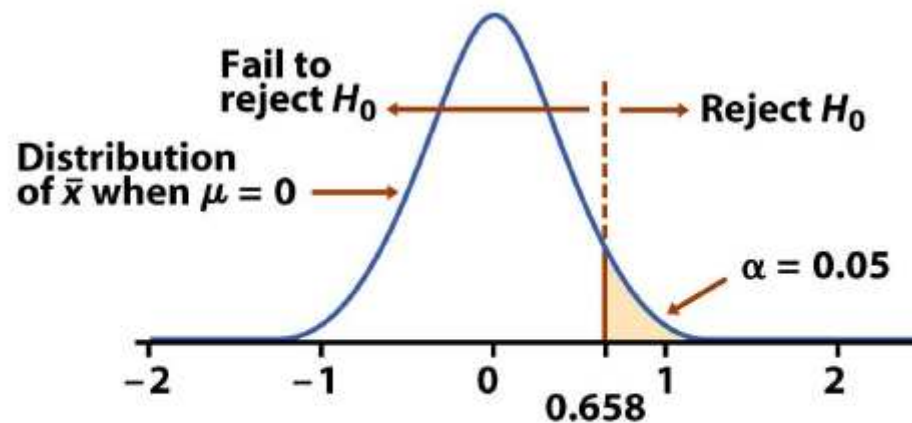
A significance level of 5% implies a lower tail of 95% and $z = 1.645$. Thus:

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

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

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

$$\bar{x} = 0.658$$



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?

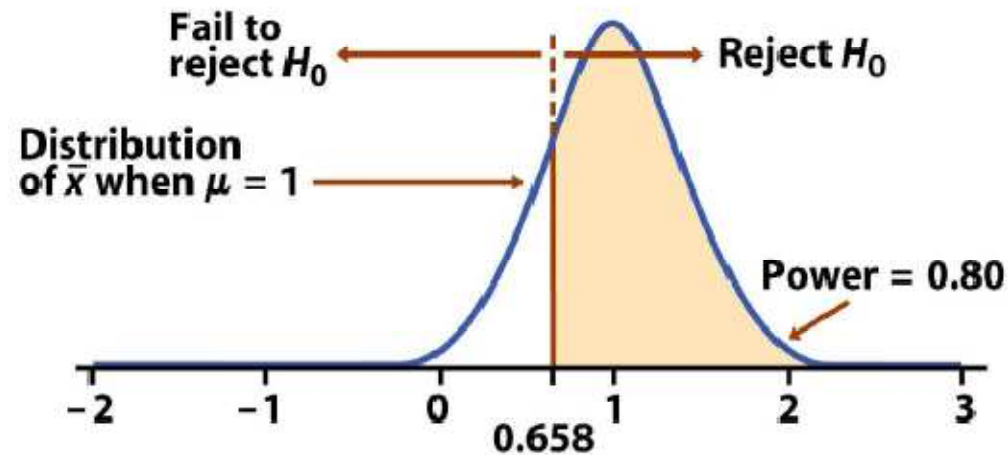
The power against the alternative $\mu = 1$ is the probability that H_0 will be rejected when in fact $\mu = 1$.

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

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

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

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.

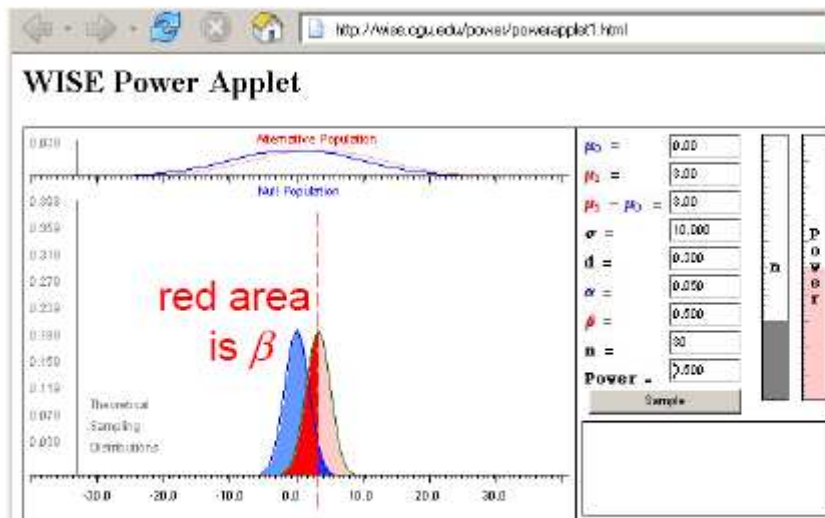
Factors affecting power: size of the effect

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 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.



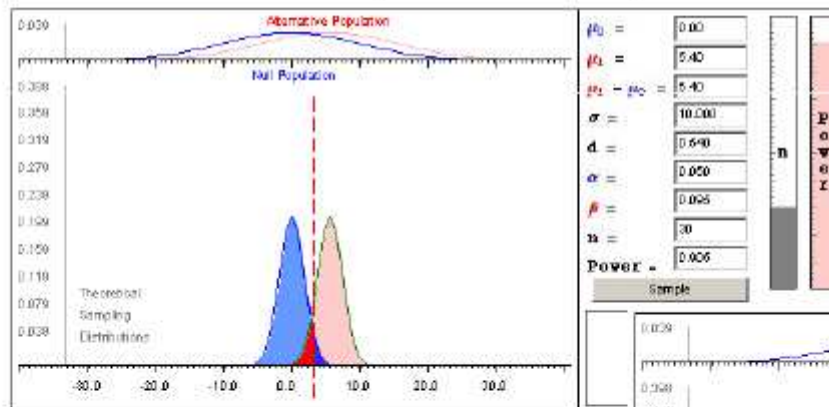
$$H_0: \mu = 0$$

$$\sigma = 10$$

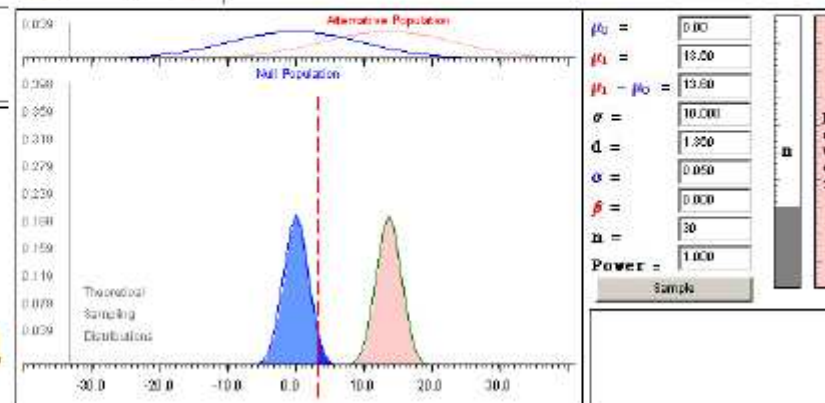
$$n = 30$$

$$\alpha = 5\%$$

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



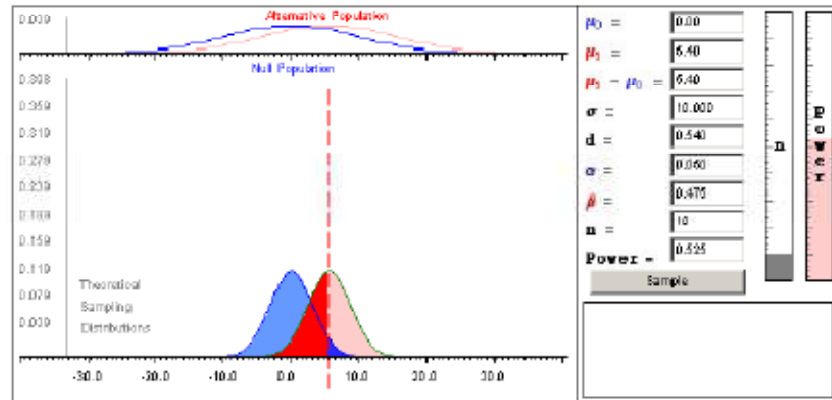
→ larger differences are easier to detect



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

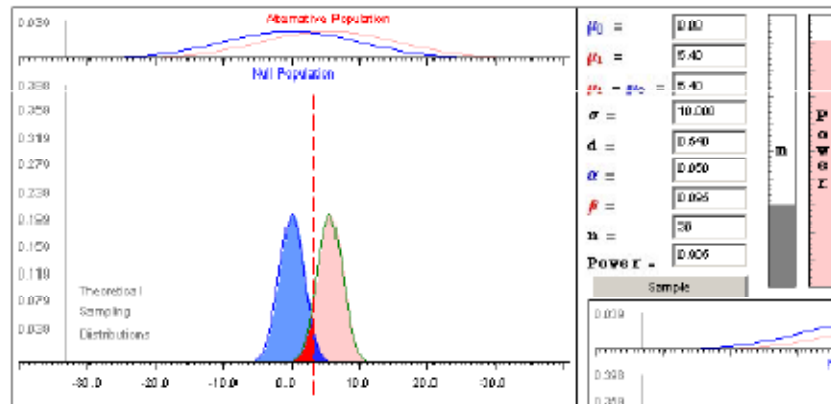


WISE Power Applet

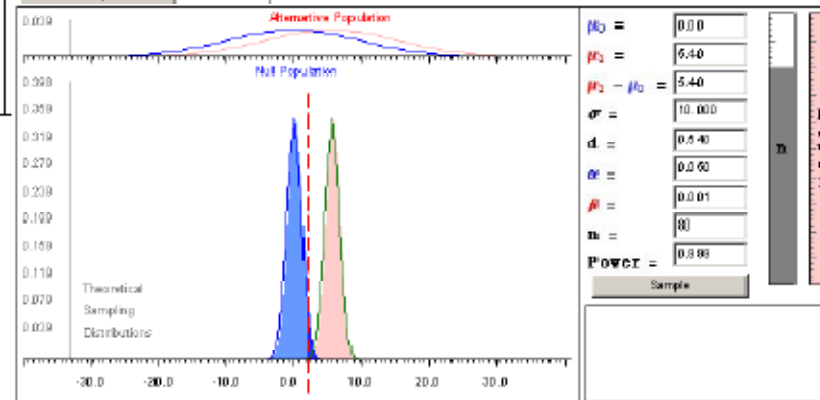


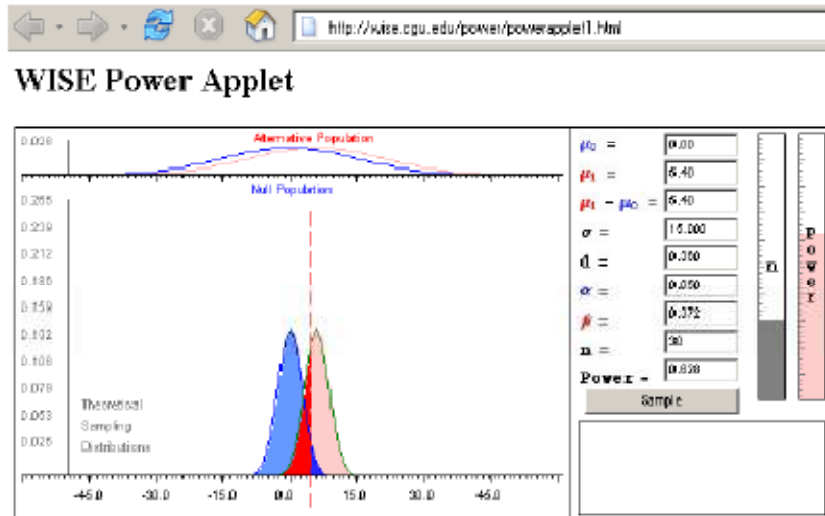
$H_0: \mu = 0$
 $\sigma = 10$
 Real $\mu = 5.4$
 $\alpha = 5\%$

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



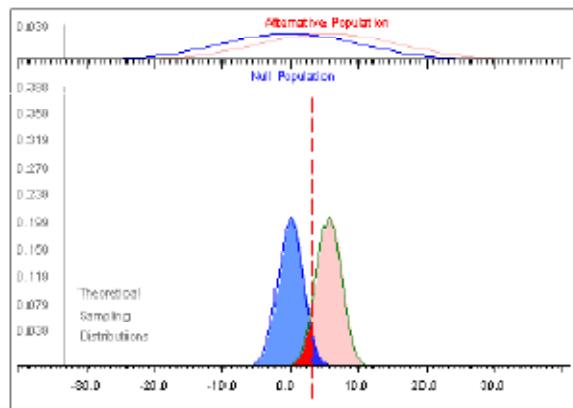
→ larger sample sizes yield greater power



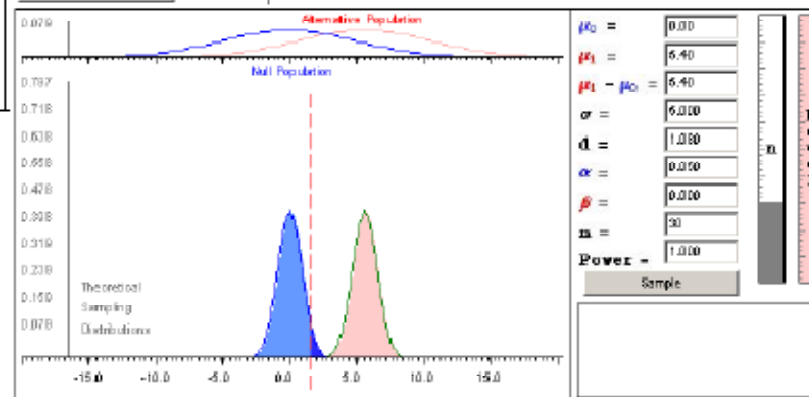


$H_0: \mu = 0$
 Real $\mu = 5.4$
 $n = 30$
 $\alpha = 5\%$

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



\rightarrow smaller variability yields greater power



Relation between power, effect size, sample size and significance level

As the effect size increases:

1. Statistical power will increase for any given sample size.
2. We need less samples to achieve a certain statistical significance.
3. The probability of Type I errors (wrong rejection) decreases.

As the sample size increases:

1. Statistical power will increase.
2. The sampling error will decrease.
3. The probability of Type II errors (not being able to reject) decreases.

As the sampling error decreases:

1. Statistical power will increase for any given effect size.
2. We need less samples to achieve the same statistical significance.
3. The probability of Type II errors (not being able to reject) decreases.

Don't panic: There exist alternatives when I cannot get more samples

Jackknife



Jackknife resampling

1. Take a subsample with all samples except 1
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 leaving out once all samples

This gives you the empirical bias and variance of the mean.

Bootstrapping: To do something seemingly impossible using only the available resources.

Bootstrap resampling

1. Take a random subsample of the original sample of size N (with replacement)
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 at least 1000 times.

This gives you the empirical distribution of the mean.

Type I and Type 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.

4.6 Multiple testing

$H_0 : p_{Head} = 0.5$



The ultimate decision is in favor of

H_0 H_1

H_0	Type I error $\alpha = 0.05$
H_1	Type II error

The true state of affairs is described by:

$p(\text{Type I error})$	$p(\text{Correct})$
α	$1 - \alpha$
α	$1 - \alpha$
$1 - (1 - \alpha)^N$	$(1 - \alpha)^N$
0.994	0.006

α → Type I error of a single test

→ Type I error of the whole family

I decide the coin is biased when it is not. Expected: 5 (=5%*100) coins

$H_0 : \mu_{drug} = \mu_{placebo} !!$

$$p(\text{Type I family error}) = 1 - (1 - \alpha)^N$$

- Choose a smaller α or equivalently, recompute the p-values of each individual test
- Fix the False Discovery Rate and choose a smaller α
- Compute empirical p-values via permutation tests or bootstrapping

Bonferroni: Choose a smaller α .

$$\alpha_{used} = \frac{\alpha_{desired}}{N} \quad \alpha_{used} = \frac{0.05}{100} = 0.0005 \quad 1 - (1 - \alpha_{used})^N = 0.0488 \quad (1 - \alpha_{used})^N = 0.9512$$

Problems: $\alpha_{used} \downarrow \Rightarrow \text{SampleSize} \uparrow$

Too conservative when tests are not independent (e.g., genes from the same person)

Acceptance of a single tests depends on the total number of tests!

The underlying assumption is that all null hypothesis are true (likely not to be true)

Low family power (not rejecting the family null hypothesis when it is false)

Recompute p-values

$$\begin{aligned}
 p_{Bonferroni} &= \min(Np_{value}, 1) \\
 p_{Bonferroni-Holm}^{(i)} &= \min((N - (i - 1))p_{value}^{(i)}, 1) \\
 p_{FDR}^{(i)} &= \min\left(\frac{N}{i} p_{value}^{(i)}, 1\right)
 \end{aligned}
 \left. \vphantom{\begin{aligned} p_{Bonferroni} \\ p_{Bonferroni-Holm}^{(i)} \\ p_{FDR}^{(i)} \end{aligned}} \right\} \text{Single-step procedures: each p-value is corrected individually.}$$

	Sequence	p-value	Bonferroni	Bonferroni-Holm	FDR
Sorted in ascending p-value ↓	1 0 1 1 1 1 1 1 1 1	0.0098	0.9800	0.0098*100=0.9800	0.0098*100/ 1=0.9800
	1 0 0 0 0 0 0 0 0 0	0.0098	0.9800	0.0098* 99=0.9668	0.0098*100/ 2=0.4883
	1 1 1 1 1 1 0 1 1 1	0.0098	0.9800	0.0098* 98=0.9570	0.0098*100/ 3=0.3255
	1 1 1 1 1 1 0 1 1 1	0.0098	0.9800	0.0098* 97=0.9473	0.0098*100/ 4=0.2441
	1 1 1 1 1 1 0 0 1 1	0.0439	1.0000	0.0439* 96=1.0000	0.0439*100/ 5=0.8789
	0 1 1 1 1 1 1 1 1 0	0.0439	1.0000	0.0439* 95=1.0000	0.0439*100/ 6=0.7324
	1 1 0 1 1 1 1 1 0 1	0.0439	1.0000	0.0439* 94=1.0000	0.0439*100/ 7=0.6278
	0 1 0 1 1 1 1 1 1 1	0.0439	1.0000	0.0439* 93=1.0000	0.0439*100/ 8=0.5493
	1 0 0 1 0 0 0 0 0 0	0.0439	1.0000	0.0439* 92=1.0000	0.0439*100/ 9=0.4883
	1 0 1 1 1 1 1 1 1 0	0.0439	1.0000	0.0439* 91=1.0000	0.0439*100/10=0.4395
	1 0 1 1 1 1 0 1 1 1	0.0439	1.0000	0.0439* 90=1.0000	0.0439*100/11=0.3995
	1 1 1 1 1 0 1 1 1 0	0.0439	1.0000	0.0439* 89=1.0000	0.0439*100/12=0.3662
	1 1 1 0 1 1 1 1 0 1	0.0439	1.0000	0.0439* 88=1.0000	0.0439*100/13=0.3380
	1 1 0 0 1 1 1 1 1 1	0.0439	1.0000	0.0439* 87=1.0000	0.0439*100/14=0.3139
	1 0 1 0 0 1 1 1 1 1	0.1172	1.0000	0.1172* 86=1.0000	0.1172*100/15=0.7813
...					

False Discovery Rate: Choose a smaller α .

$$P(H_0 | p \leq \alpha) = \frac{P(p \leq \alpha | H_0)P(H_0)}{P(p \leq \alpha | H_0)P(H_0) + P(p \leq \alpha | H_1)P(H_1)} = \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Proportion of tests that follows the null hypothesis

$\alpha\pi_0$ $\alpha\pi_0$

$\alpha\pi_0$ $(1-\beta)$ $(1-\pi_0)$

Proportion of false positives Power to detect tests following the alternative hypothesis Proportion of tests following the alternative hypothesis

$$\alpha_{used} = \underbrace{\frac{P(H_0 | p \leq \alpha)}{1 - P(H_0 | p \leq \alpha)}}_{\text{FDR}} \frac{1 - \pi_0}{\pi_0} (1 - \beta)$$

Example: FDR=0.05, $\pi_0 = 0.9$, $\beta = 0.3$

$$\alpha_{used} = 0.05 \frac{1 - 0.9}{0.9} (1 - 0.3) = 0.0039$$

- What about prefiltering experiments (according to intensity, variance etc.) to reduce the proportion of false positives - e.g. tests with consistently low intensity may not be considered interesting?
- Can be useful, but:
 - The criteria for filtering have to be chosen before the analysis, i.e. not dependent on the results of the analysis.
 - The criteria have to be independent of the distribution of the test statistic under the null hypothesis - otherwise no control of the type I error.

5 Inference for distributions

5.1 Inference for the mean of a population

Introduction

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:

$$H_0: \mu = 0 \text{ versus } H_a: \mu > 0$$

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.

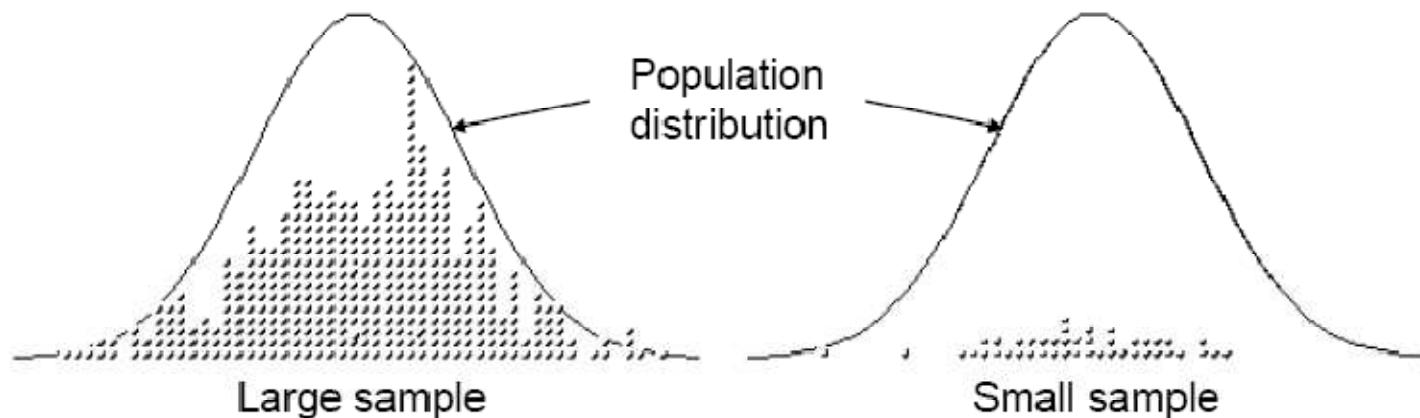
This situation is very common with real data.

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 mediocre estimate of σ .



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

For a sample of size n ,
the sample standard deviation s is:
 $n - 1$ is the “degrees of freedom.”

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

The value s/\sqrt{n} is called the standard error of the mean **SEM**.

Scientists often present sample results as mean \pm SEM.



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

What is the standard deviation s of the sample data?

$$\begin{aligned} \text{SEM} = s/\sqrt{n} &\Leftrightarrow s = \text{SEM} \cdot \sqrt{n} \\ s &= 8.9 \cdot \sqrt{25} = 44.5 \end{aligned}$$

The t-distribution

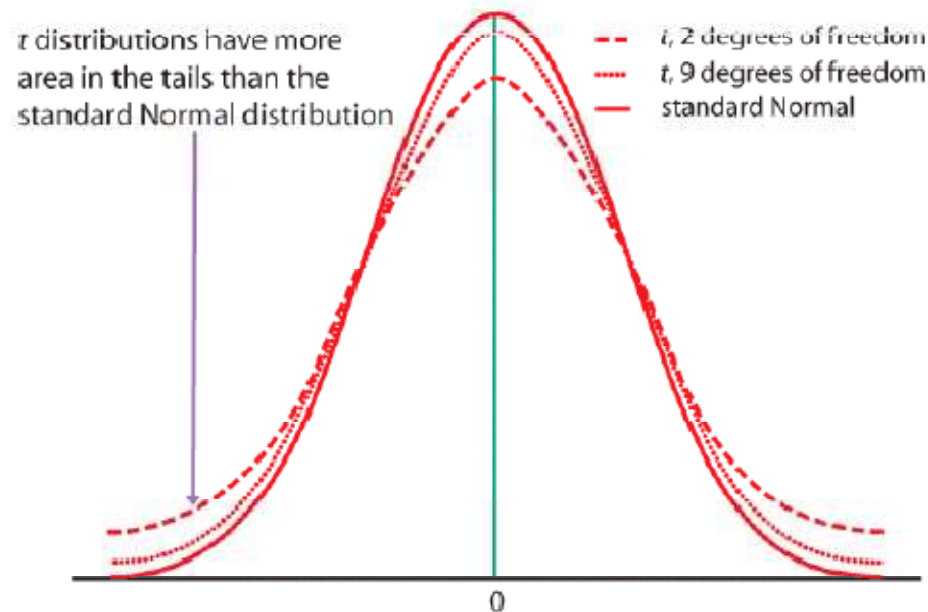
Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population.

- 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 $n - 1$** .

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

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.

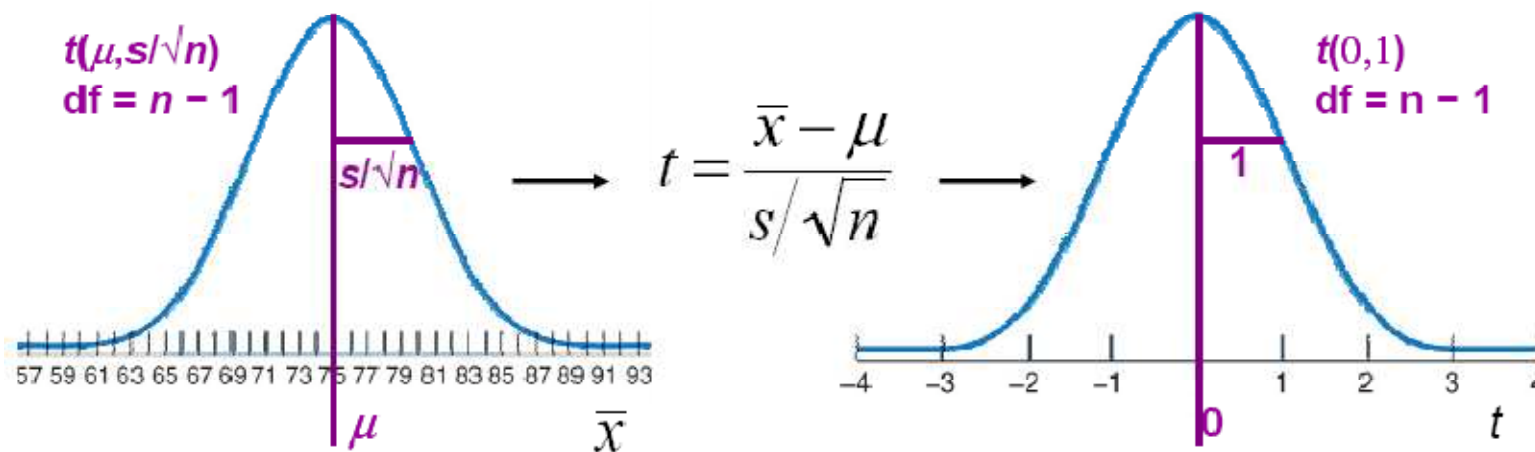
The t distributions become wider for smaller sample sizes, reflecting the lack of precision in estimating σ from s .



Standardizing data before using Table D

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

Then we can use **Table D** 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/\sqrt{n} is its standard deviation (width).
You obtain s , the standard deviation of the sample, with your calculator.

Table D

When σ is unknown, we use a t distribution with “ $n-1$ ” degrees of freedom (df).

Table D shows the z -values and t -values corresponding to landmark P -values/ confidence levels.

When σ is known, we use the normal distribution and the standardized z -value.

df	Upper tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.764	2.150	2.272	2.640	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.759	2.144	2.266	2.634	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.755	2.139	2.261	2.629	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.751	2.134	2.257	2.624	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.748	2.130	2.253	2.620	2.876	3.197	3.611	3.922
19	0.688	0.861	1.066	1.327	1.745	2.126	2.250	2.616	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.324	1.742	2.122	2.247	2.612	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.740	2.120	2.245	2.610	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.738	2.118	2.243	2.608	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.067	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

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

Table A vs. Table D

Table A gives the area to the LEFT of hundreds of z-values.

It should only be used for Normal distributions.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

(...)

Table D t distribution critical values

df	Upper tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
<i>z*</i>	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 <i>C</i>											

Table D gives the area to the RIGHT of a dozen *t* or *z*-values.

It can be used for *t* distributions of a given *df* and for the Normal distribution.

Table D also gives the middle area under a *t* or normal distribution comprised between the negative and positive value of *t* or *z*.

The one-sample t-confidence interval

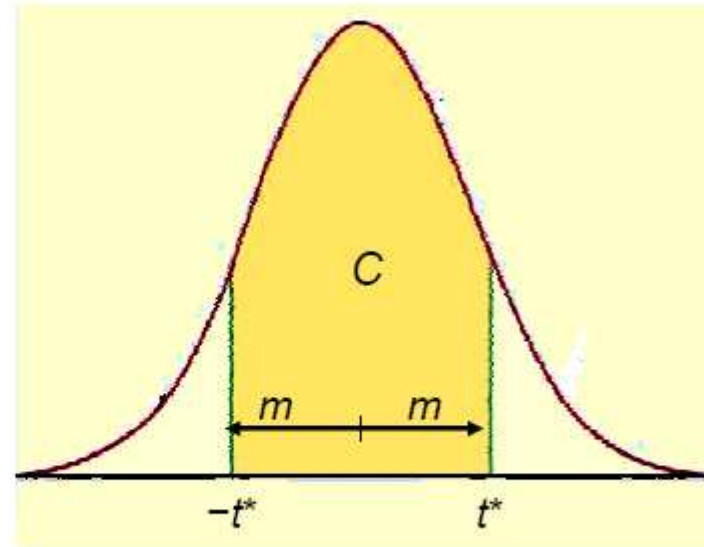
The **level C confidence interval** is an interval with probability C of containing the true population parameter.

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 D for $df = n-1$ and confidence level C .
- The margin of error m is:

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



Red wine, in moderation

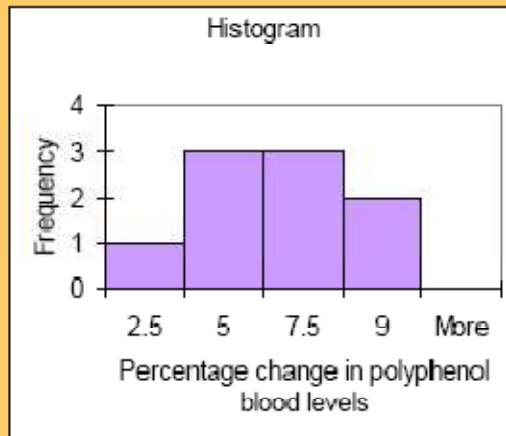


Drinking red wine in moderation may protect against heart attacks. The polyphenols it contains act on blood cholesterol, likely helping to prevent heart attacks.

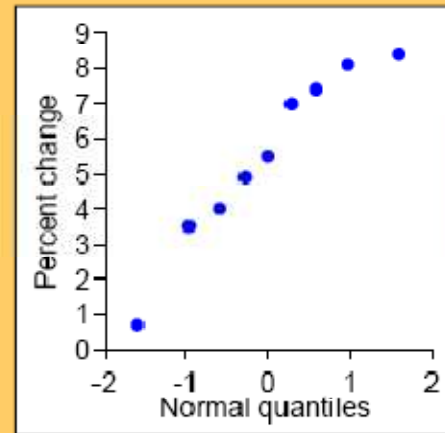
To see if moderate red wine consumption increases the average blood level of polyphenols, a group of nine randomly selected healthy men were assigned to drink half a bottle of red wine daily for two weeks. Their blood polyphenol levels were assessed before and after the study, and the percent change 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.

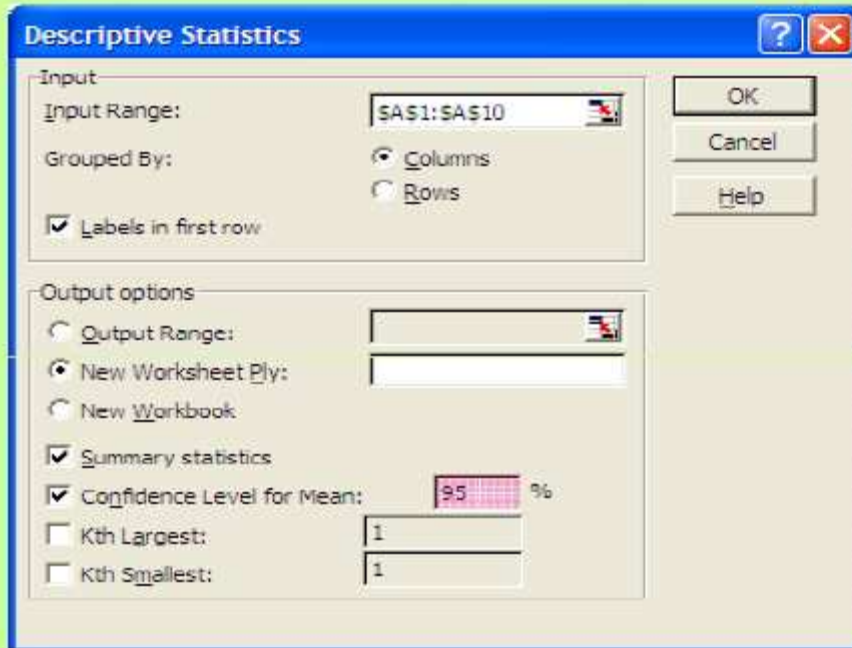
For $df = 8$ and $C = 95\%$, $t^* = 2.306$.

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

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.

Excel

Menu: Tools/DataAnalysis: select "Descriptive statistics"



PercentChange	
Mean	5.5
Standard Error	0.838981 s/\sqrt{n}
Median	5.5
Mode	#N/A
Standard Deviation	2.516943
Sample Variance	6.335
Kurtosis	0.010884
Skewness	-0.7054
Range	7.7
Minimum	0.7
Maximum	8.4
Sum	49.5
Count	9
Confidence Level(95.0%)	1.934695 m

!!! Warning: do not use the function =CONFIDENCE(alpha, stdev, size)

This assumes a normal sampling distribution (stdev here refers to σ)

and uses z^* instead of t^* !!!

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 D
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 :

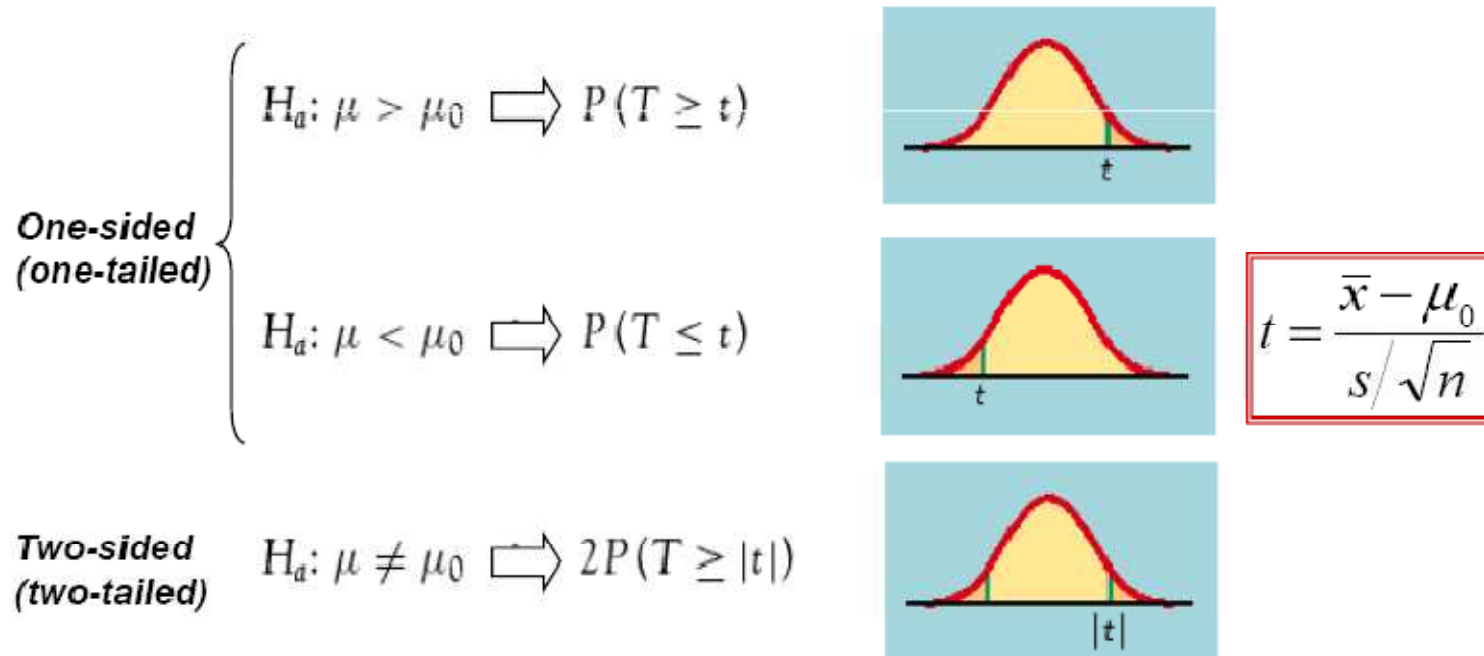
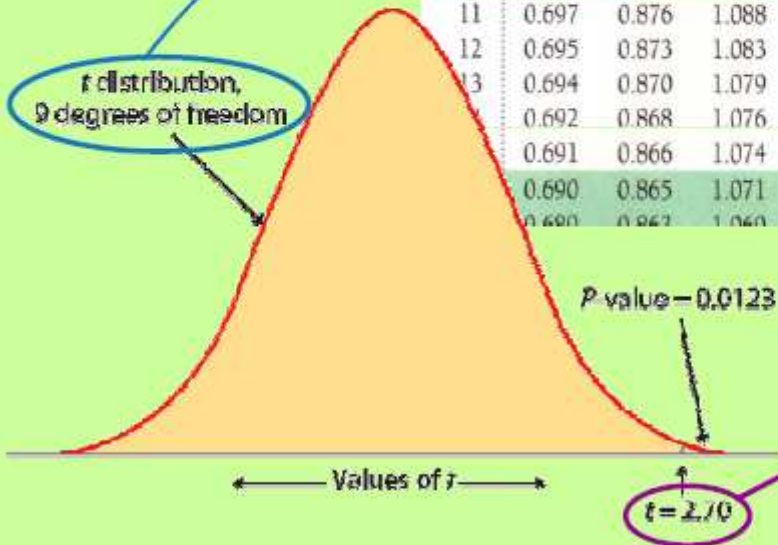


Table D

df	Upper tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.221	2.567	2.898	3.223	3.646	3.966

For $df = 9$ we only look into the corresponding row.

t distribution, 9 degrees of freedom



The calculated value of t is 2.7. We find the 2 closest t values.

$$2.398 < t = 2.7 < 2.821$$

thus

$$0.02 > \text{upper tail } p > 0.01$$

For a one-sided H_a , this is the P-value (between 0.01 and 0.02);
 for a two-sided H_a , the P-value is doubled (between 0.02 and 0.04).

Excel

TDIST(x, degrees_freedom, tails)

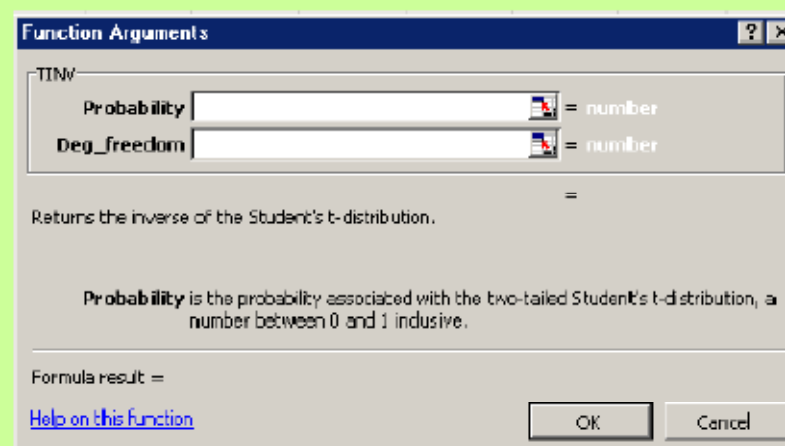
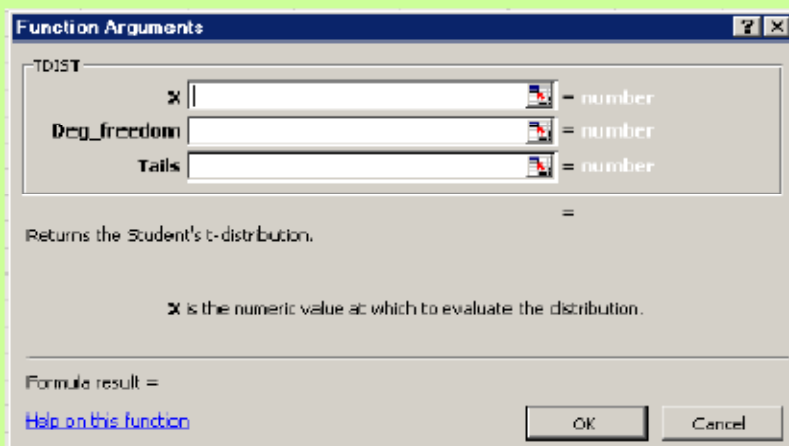
TDIST = $P(X > x)$ for a random variable X following the t distribution (x positive).
Use it in place of Table C or to obtain the p-value for a positive t-value.

- ❑ X is the *standardized* value at which to evaluate the distribution (i.e., “t”).
- ❑ *Degrees_freedom* is an integer indicating the number of degrees of freedom.
- ❑ *Tails* specifies the number of distribution tails to return. If tails = 1, TDIST returns the one-tailed p-value. If tails = 2, TDIST returns the two-tailed p-value.

TINV(probability,degrees_freedom)

Gives the t-value (e.g., t^*) for a given probability and degrees of freedom.

- ❑ *Probability* is the probability associated with the two-tailed t distribution.
- ❑ *Degrees_freedom* is the number of degrees of freedom of the t distribution.



Sweetening colas (continued)

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



$H_0: \mu = 0$ 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 $t_\alpha = 1.833$.
 $t > t_\alpha$ 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.

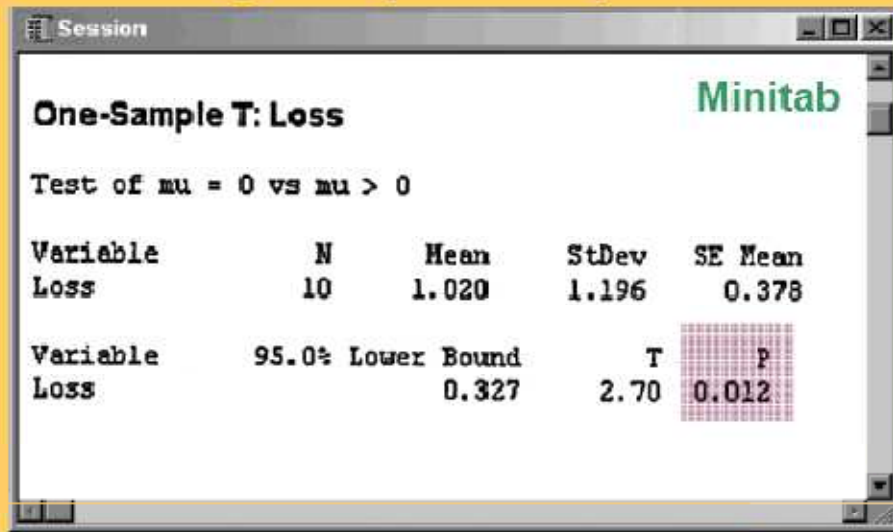
Taster	Sweetness loss
1	2.0
2	0.4
3	0.7
4	2.0
5	-0.4
6	2.2
7	-1.3
8	1.2
9	1.1
10	2.3
<hr/>	
Average	1.02
Standard deviation	1.196
Degrees of freedom	$n - 1 = 9$

df	Upper tail probability p											
	.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.

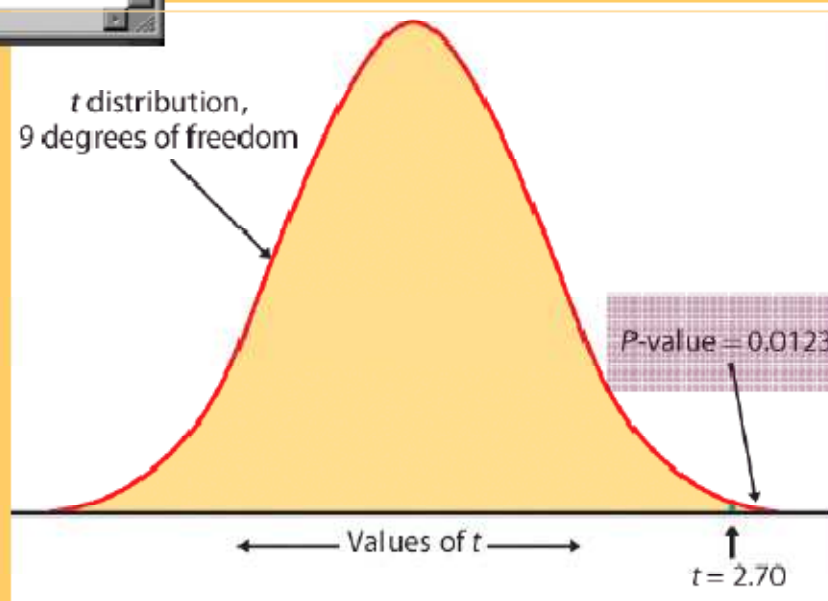
Sweetening colas (continued)



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

$$df = n - 1 = 9$$

In Excel, you can obtain the precise p-value once you have calculated t .
 Use the function `dist(t , df , tails)`
 “=tdist(2.7, 9, 1),” which gives 0.01226



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 $X_{\text{difference}} = (X_1 - X_2)$, and

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

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:

$$H_0: \mu = 0 \text{ versus } H_a: \mu > 0$$

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.

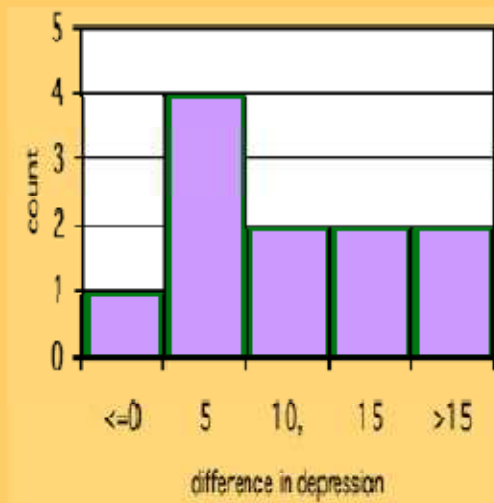
A matched pairs test of significance is indeed just like a one-sample test.

Does lack of caffeine increase depression?

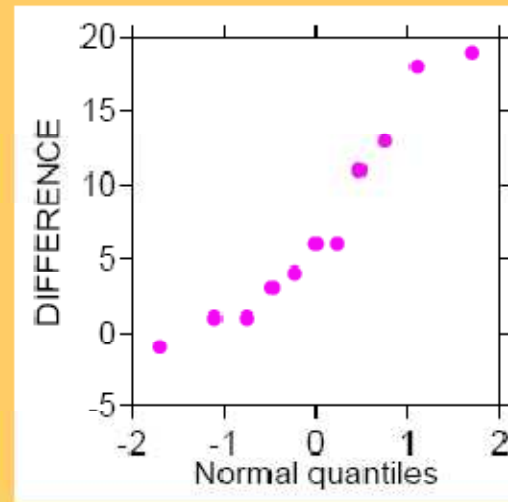
Individuals diagnosed as caffeine-dependent are deprived of caffeine-rich foods and assigned to receive daily pills. Sometimes, the pills contain caffeine and other times they contain a placebo. Depression was assessed.

Subject	Depression with Caffeine	Depression with Placebo	Placebo - Caffeine
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 $\bar{x} = 7.36$; $s = 6.92$

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

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

Subject	Depression with Caffeine	Depression with Placebo	Placebo - Caffeine
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

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

Caffeine deprivation causes a significant increase in depression.

SPSS statistical output for the caffeine study:

- Conducting a paired sample t -test on the raw data (caffeine and placebo)
- Conducting a one-sample t -test on difference (caffeine – placebo)

Paired Samples Test								
	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Placebo - Caffeine	7.364	6.918	2.086	2.716	12.011	3.530	10	.005

One-Sample Test						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Difference	3.530	10	.005	7.364	2.72	12.01

Our alternative hypothesis was one-sided, thus our p-value is half of the two-tailed p-value provided in the software output (half of 0.005 = 0.0025).

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 for 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 value of the mean 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 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 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

$t(\text{df}:15) = 1.729$. 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:

$$\begin{aligned} P(\bar{x} \geq 1.068 \text{ when } \mu = 3) &= P\left(z \geq \frac{1.068 - 3}{7/\sqrt{16}}\right) \\ &= P(z \geq -1.10) = 1 - P(z \leq -1.10) = 0.8643 \end{aligned}$$

The power would be about 86%.

Inference for non-normal distributions

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

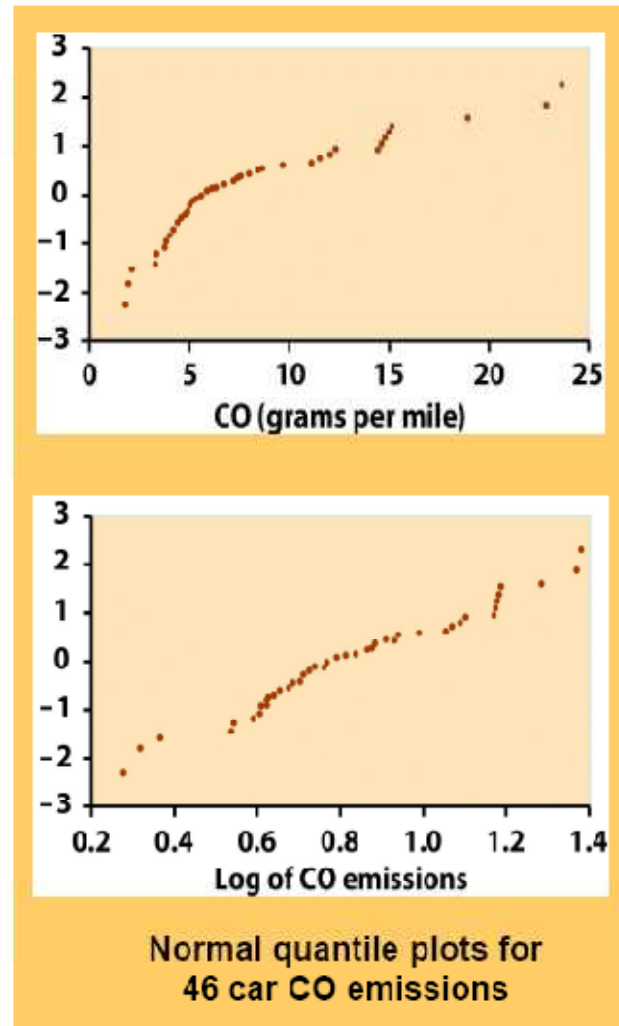
- ❑ If the data are skewed, you can attempt to **transform** the variable to bring it closer to normality (e.g., logarithm transformation). The t -procedures applied to transformed data are quite accurate for even moderate sample sizes.
- ❑ A distribution other than a normal distribution might describe 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 (see chapter 15) 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.



Non-parametric 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.

$$H_0: \text{population median} = 0 \quad \text{vs.} \quad H_a: \text{population median} > 0$$

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.

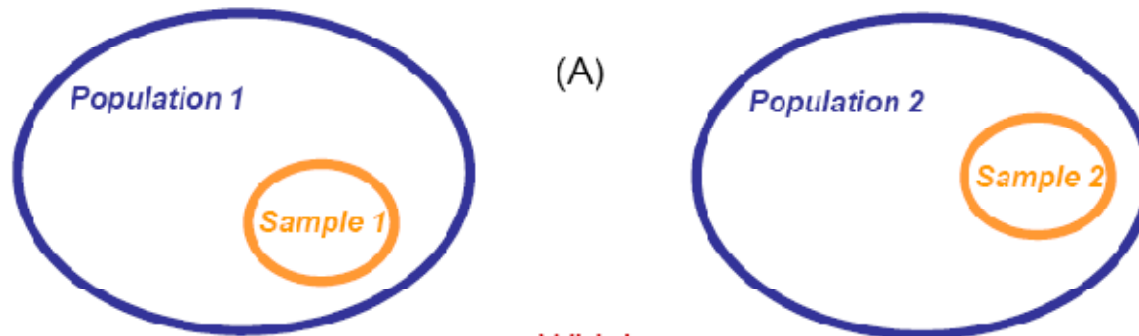
The test statistic is the count X of pairs with a positive difference.

P-values for X are based on the binomial $B(n, 1/2)$ distribution.

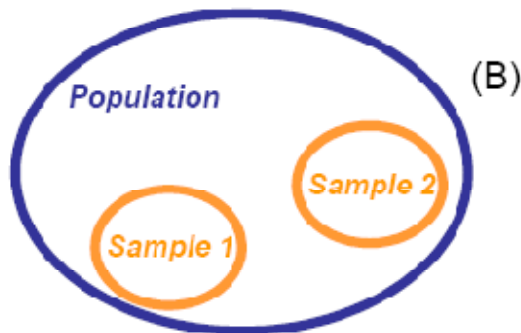
$$H_0: p = 1/2 \quad \text{vs.} \quad H_a: p > 1/2$$

5.2 Comparing two means

Comparing two samples



Which
is it?



We often compare two treatments used on **independent** samples.

Is the difference between both treatments due only to variations from the random sampling (B), or does it reflect a true difference in population means (A)?

Independent samples: Subjects in one samples are completely unrelated to subjects in the other sample.

Two-sample z-statistic

We have **two independent SRSs** (simple random samples) possibly coming from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2) . We use \bar{x}_1 and \bar{x}_2 to estimate the unknown μ_1 and μ_2 .

When both populations are normal, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$

is also normal, with standard deviation :

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Then the **two-sample z statistic** has the standard normal $N(0, 1)$ sampling distribution.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Two independent samples t-distribution

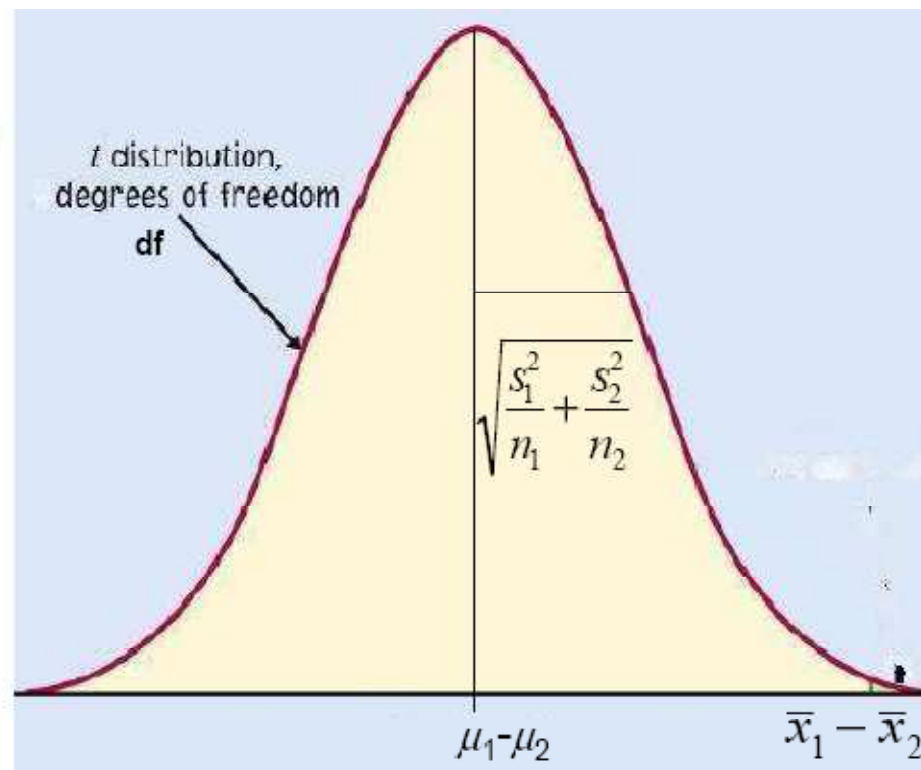
We have **two independent SRSs** (simple random samples) possibly coming from two distinct populations with (μ_1, σ_1) and (μ_2, σ_2) unknown. We use (\bar{x}_1, s_1) and (\bar{x}_2, s_2) to estimate (μ_1, σ_1) and (μ_2, σ_2) , respectively.

To compare the means, both populations should be normally distributed. However, in practice, it is enough that the two distributions have similar shapes and that the sample data contain no strong outliers.

The two-sample t statistic follows approximately the t distribution with a standard error SE (spread) reflecting variation from both samples:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Conservatively, the degrees of freedom is equal to the smallest of $(n_1 - 1, n_2 - 1)$.



Two-sample t-significance test

The null hypothesis is that both population means μ_1 and μ_2 are equal, thus their difference is equal to zero.

$$H_0: \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$$

with either a one-sided or a two-sided alternative hypothesis.

We find how many standard errors (SE) away from $(\mu_1 - \mu_2)$ is $(\bar{x}_1 - \bar{x}_2)$ by standardizing with t :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

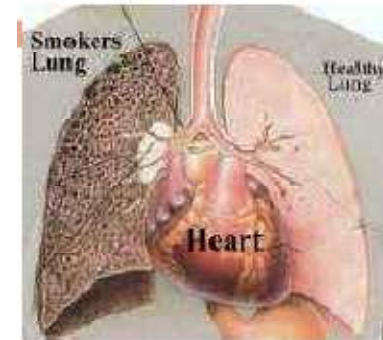
Because in a two-sample test H_0 poses $(\mu_1 - \mu_2) = 0$, we simply use

With $df = \text{smallest}(n_1 - 1, n_2 - 1)$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Does smoking damage the lungs of children exposed to parental smoking?

Forced vital capacity (FVC) is the volume (in milliliters) of air that an individual can exhale in 6 seconds.



FVC was obtained for a sample of children not exposed to parental smoking and a group of children exposed to parental smoking.

Parental smoking	FVC \bar{x}	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

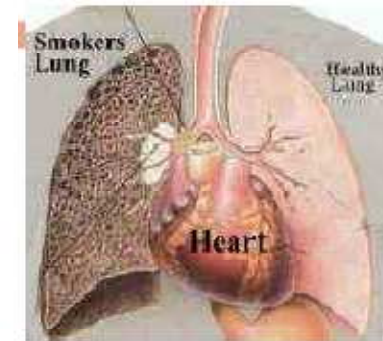


We want to know whether parental smoking decreases children's lung capacity as measured by the FVC test.

Is the mean FVC lower in the population of children exposed to parental smoking?

$$H_0: \mu_{\text{smoke}} = \mu_{\text{no}} \Leftrightarrow (\mu_{\text{smoke}} - \mu_{\text{no}}) = 0$$

$$H_a: \mu_{\text{smoke}} < \mu_{\text{no}} \Leftrightarrow (\mu_{\text{smoke}} - \mu_{\text{no}}) < 0 \text{ (one sided)}$$



The difference in sample averages follows approximately the t distribution: $t \left(0, \sqrt{\frac{S_{\text{smoke}}^2}{n_{\text{smoke}}} + \frac{S_{\text{no}}^2}{n_{\text{no}}}} \right), df \ 29$

We calculate the t statistic:

$$t = \frac{\bar{x}_{\text{smoke}} - \bar{x}_{\text{no}}}{\sqrt{\frac{S_{\text{smoke}}^2}{n_{\text{smoke}}} + \frac{S_{\text{no}}^2}{n_{\text{no}}}}} = \frac{75.5 - 88.2}{\sqrt{\frac{9.3^2}{30} + \frac{15.1^2}{30}}}$$

Parental smoking	FVC \bar{x}	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

$$t = \frac{-12.7}{\sqrt{2.9 + 7.6}} \approx -3.9$$

In table D, for $df \ 29$ we find:

$$|t| > 3.659 \Rightarrow p < 0.0005 \text{ (one sided)}$$

It's a very significant difference, we reject H_0 .

Lung capacity is significantly impaired in children of smoking parents.

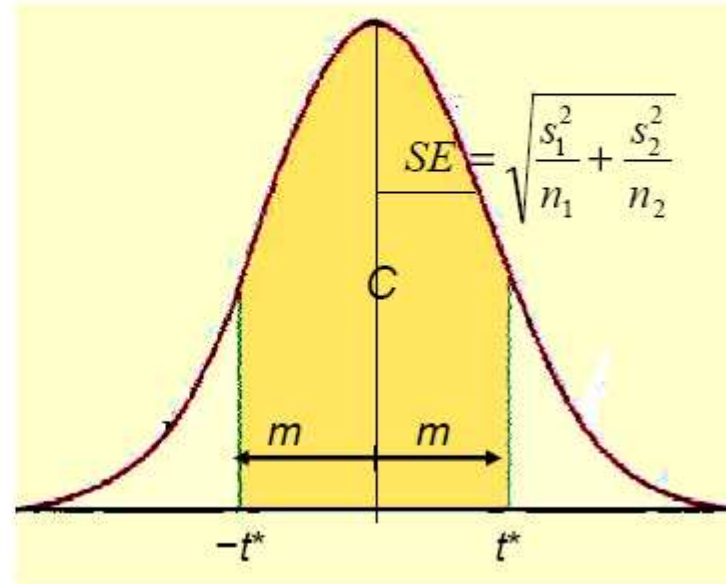
Two-sample t-confidence interval

Because we have two independent samples we use the difference between both sample averages ($\bar{x}_1 - \bar{x}_2$) to estimate $(\mu_1 - \mu_2)$.

Practical use of t^*

- ▣ C is the area between $-t^*$ and t^* .
- ▣ We find t^* in the line of Table D for $df = \text{smallest}(n_1 - 1; n_2 - 1)$ and the column for confidence level C .
- ▣ The margin of error m is:

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = t^* SE$$



Common mistake

A common mistake is to calculate a one-sample confidence interval for μ_1 and then check whether μ_2 falls within that confidence interval, or vice-versa.

This is WRONG because the variability in the sampling distribution for two independent samples is more complex and must take into account variability coming from both samples. Hence the more complex formula for the standard error.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Can directed reading activities in the classroom help improve reading ability? A class of 21 third-graders participates in these activities for 8 weeks while a control classroom of 23 third-graders follows the same curriculum without the activities. After 8 weeks, all children take a reading test (scores in table).

Treatment group				Control group				Group	<i>n</i>	\bar{x}	<i>s</i>
24	61	59	46	42	33	46	37	Treatment	21	51.48	11.01
43	44	52	43	43	41	10	42	Control	23	41.52	17.15
58	67	62	57	55	19	17	55				
71	49	54		26	54	60	28				
43	53	57		62	20	53	48				
49	56	33		37	85	42					

95% confidence interval for $(\mu_1 - \mu_2)$, with $df = 20$ conservatively $\rightarrow t^* = 2.086$:

$$CI : (\bar{x}_1 - \bar{x}_2) \pm m; \quad m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.086 * 4.31 \approx 8.99$$

With 95% confidence, $(\mu_1 - \mu_2)$, falls within 9.96 ± 8.99 or 1.0 to 18.9.

20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Robustness

The two-sample t procedures are more robust than the one-sample t procedures. They are the most robust when both sample sizes are equal and both sample distributions are similar. But even when we deviate from this, two-sample tests tend to remain quite robust.

- ➔ When planning a two-sample study, choose equal sample sizes if you can.

As a guideline, a combined sample size ($n_1 + n_2$) of 40 or more will allow you to work with even the most skewed distributions.

Details of the two sample t procedures

The **true value of the degrees of freedom** for a two-sample t -distribution is quite lengthy to calculate. That's why we use an approximate value, $df = \text{smallest}(n_1 - 1, n_2 - 1)$, which errs on the conservative side (often smaller than the exact).

Computer software, though, gives the exact degrees of freedom—or the rounded value—for your sample data.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

95% confidence interval for the reading ability study using the more precise degrees of freedom:

$$df = \frac{\left(\frac{11.01^2}{21} + \frac{17.15^2}{23}\right)^2}{\frac{1}{20} \left(\frac{11.01^2}{21}\right)^2 + \frac{1}{22} \left(\frac{17.15^2}{23}\right)^2}$$

$$= \frac{344.486}{9.099} = 37.86$$

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$m = 2.024 * 4.31 \approx 8.72$$

30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423
	50%	60%	70%	80%	90%	95%	96%	98%
	Confidence level C							

t-Test: Two-Sample Assuming Unequal Variances

Excel

	Treatment group	Control group
Mean	51.476	41.522
Variance	121.162	294.079
Observations	21	23
Hypothesized Mean Difference	-	
df	38	
t Stat	2.311	
P(T<=t) one-tail	0.013	
t Critical one-tail	1.686	
P(T<=t) two-tail	0.026	
t Critical two-tail	2.024	t*

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Reading Score	Equal variances assumed	2.362	.132	2.267	42	.029	9.95445	4.39189	1.09125	18.81765
	Equal variances not assumed			2.311	37.855	.026	9.95445	4.30763	1.23302	18.67588

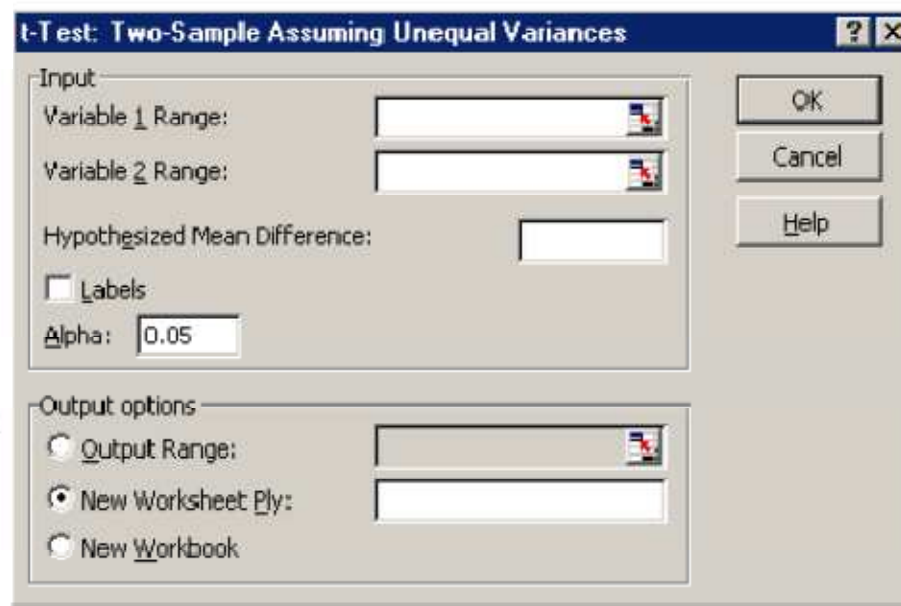
Excel

menu/tools/data_analysis →

or

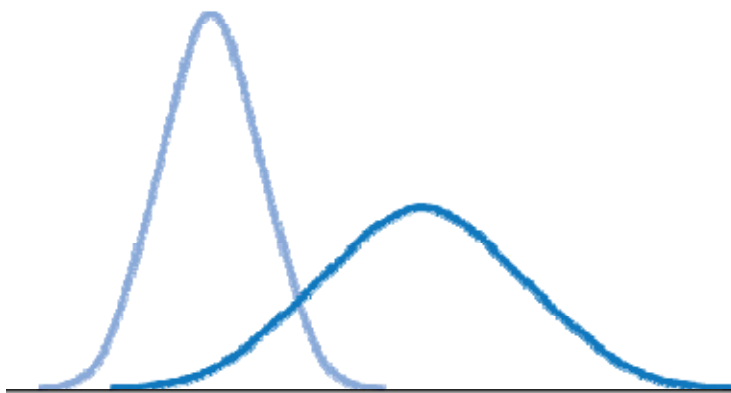
=TTEST(array1,array2,tails,type)

- ❑ *Array1* is the first data set.
- ❑ *Array2* is the second data set.
- ❑ *Tails* specifies the nature of the alternative hypothesis
(1: one-tailed; 2: two-tailed).
- ❑ *Type* is the kind of *t*-test to perform
(1: paired; 2: two-sample equal variance; 3: two-sample unequal variance).



Pooled two-sample procedures

There are two versions of the two-sample t -test: one **assuming equal variance** (“pooled 2-sample test”) and one **not assuming equal variance** (“unequal” variance, as we have studied) for the two populations. They have slightly different formulas and degrees of freedom.



Two normally distributed populations with unequal variances

The pooled (equal variance) two-sample t -test was often used before computers because it has exactly the t distribution for degrees of freedom $n_1 + n_2 - 2$.

However, the assumption of equal variance is hard to check, and thus the unequal variance test is safer.

When both population have the *same* standard deviation, the **pooled estimator of σ^2** is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The sampling distribution for $(x_1 - x_2)$ has exactly the t distribution with **$(n_1 + n_2 - 2)$ degrees of freedom.**

A level C confidence interval for $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
(with area C between $-t^*$ and t^*)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

To test the hypothesis $H_0: \mu_1 = \mu_2$ against a one-sided or a two-sided alternative, compute the pooled two-sample t statistic for the $t(n_1 + n_2 - 2)$ distribution.

Which type of test? One sample, two samples, paired samples?

- Comparing vitamin content of bread immediately after baking vs. 3 days later (the same loaves are used on day one and 3 days later).
- Comparing vitamin content of bread immediately after baking vs. 3 days later (tests made on independent loaves).
- Average fuel efficiency for 2005 vehicles is 21 miles per gallon. Is average fuel efficiency higher in the new generation “green vehicles”?
- Is blood pressure altered by use of an oral contraceptive? Comparing a group of women not using an oral contraceptive with a group taking it.
- Review insurance records for dollar amount paid after fire damage in houses equipped with a fire extinguisher vs. houses without one. Was there a difference in the average dollar amount paid?

5.3 Optional topics in comparing distributions

Inference for population spread

It is also possible to compare two population standard deviations σ_1 and σ_2 by comparing the standard deviations of two SRSs. However, these procedures are **not robust at all against deviations from normality**.

When s_1^2 and s_2^2 are sample variances from independent SRSs of sizes n_1 and n_2 drawn from normal populations, the F statistic

$$F = s_1^2 / s_2^2$$

has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when $H_0: \sigma_1 = \sigma_2$ is true.

The F distributions are right-skewed and cannot take negative values.

- The peak of the F density curve is near 1 when both population standard deviations are equal.
- Values of F far from 1 in either direction provide evidence against the hypothesis of equal standard deviations.

Table E in the back of the book gives critical F -values for upper p of 0.10, 0.05, 0.025, 0.01, and 0.001. We compare the F statistic calculated from our data set with these critical values for a one-side alternative; the p -value is doubled for a two-sided alternative.

$$F \text{ has } \frac{Df_{\text{numerator}} : n_1 - 1}{Df_{\text{denom}} : n_2 - 1}$$

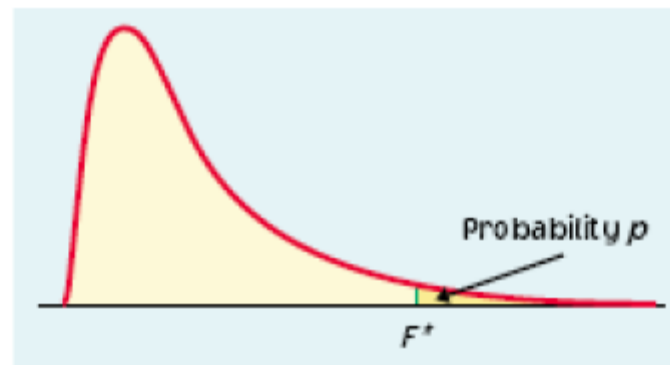


Table E F distribution critical values

$df_{num} = n_1 - 1$

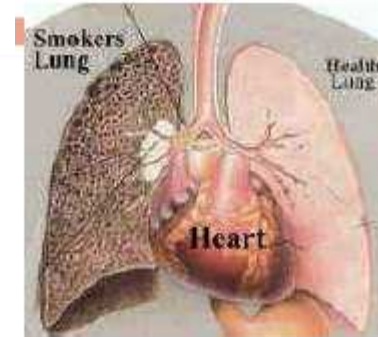
		Degrees of freedom in the numerator							
		1	2	3	4	5	6	7	8
p									
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
	0.025	647.79	799.50	864.16	899.58	928.85	937.11	948.22	956.66
	0.010	4052.2	4999.5	5403.4	5624.6	5768.6	5859	5928.4	5981.1
	0.001	405284	500000	540379	562500	576405	585937	592873	598144
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98
	0.010	21.20	18.60	16.69	15.98	15.52	15.21	14.98	14.80
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03

F

$df_{den} = n_2 - 1$

Does parental smoking damage the lungs of children?

Forced vital capacity (FVC) was obtained for a sample of children not exposed to parental smoking and a group of children exposed to parental smoking.



Parental smoking	FVC \bar{x}	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

$$H_0: \sigma^2_{\text{smoke}} = \sigma^2_{\text{no}}$$

$$H_a: \sigma^2_{\text{smoke}} \neq \sigma^2_{\text{no}} \text{ (two sided)}$$

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2} = \frac{15.1^2}{9.3^2} \approx 2.64$$

The degrees of freedom are 29 and 29, which can be rounded to the closest values in Table E: 30 for the numerator and 25 for the denominator.

$$2.54 < F(30,25) = 2.64 < 3.52$$

$$\rightarrow 0.01 > 1\text{-sided } p > 0.001$$

$$\rightarrow 0.02 > 2\text{-sided } p > 0.002$$

F [*]	Proba	Degrees of freedom (Df) in the numerator												
		1	2	3	4	5	6	7	8	9	10	15	20	30
25	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.77	1.72	1.66
	0.050	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.09	2.01	1.92
	0.025	5.69	4.29	3.69	3.36	3.13	2.97	2.85	2.76	2.68	2.61	2.41	2.3	2.18
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.7	2.54
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.06	3.79	3.52
50	0.100	2.81	2.41	2.2	2.06	1.97	1.9	1.84	1.8	1.76	1.73	1.63	1.57	1.5
	0.050	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07	2.03	1.87	1.78	1.69
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.99	1.87
	0.010	7.17	5.06	4.2	3.72	3.41	3.19	3.02	2.89	2.78	2.7	2.42	2.27	2.1
	0.001	12.22	7.96	6.34	5.46	4.9	4.51	4.22	4	3.82	3.67	3.2	2.95	2.66

Power of two-sample t -test

The power of the two-sample t -test for a specific alternative value of the difference in population means $(\mu_1 - \mu_2)$, assuming a fixed significance level α , is the probability that the test will reject the null hypothesis when the alternative is true.

The basic concept is similar to that for the one-sample t -test. The exact method involves the **noncentral t distribution**. Calculations are carried out with software.

You need information from a pilot study or previous research to calculate an expected power for your t -test and this allows you to plan your study smartly.

Power calculations using a non-central t-distribution

For the pooled two-sample t -test, with parameters μ_1 , μ_2 , and the common standard deviation σ we need to specify:

- An alternative that would be important to detect (i.e., a value for $\mu_1 - \mu_2$)
- The sample sizes, n_1 and n_2
- The Type I error for a fixed significance level, α
- A guess for the standard deviation σ

We find the degrees of freedom $df = n_1 + n_2 - 2$ and the value of t^* that will lead to rejection of $H_0: \mu_1 - \mu_2 = 0$

Then we calculate the **non-centrality parameter** δ

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$


Lastly, we find the power as the probability that a noncentral t random variable with degrees of freedom df and noncentrality parameter δ will be less than t^* :


- In SAS this is $1 - \text{PROBT}(t^*, df, \delta)$. There are also several free online tools that calculate power.
- Without access to software, we can approximate the power as the probability that a standard normal random variable is greater than $t^* - \delta$, that is, $P(z > t^* - \delta)$, and use Table A.


For a ***test with unequal variances*** we can simply use the conservative degrees of freedom, but we need to guess both standard deviations and combine them for the guessed standard error:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Online tools:

 <http://www.stat.uiowa.edu/~rlenth/Power/>

 Normal Power Calculations

 Russ Lenth's power and sample-size ...



Java applets for power and sample size

Select the analysis to be used in your study

- CI for one proportion
- Test of one proportion
- Test comparing two proportions
- CI for one mean
- One-sample t test (or paired t)
- Two-sample t test (pooled or Satterthwaite)**
- Linear regression
- Balanced ANOVA (any model)
- Generic chi-square test
- Generic Poisson test

Run

This software is intended to be useful in planning statistical studies. It is not intended to be used for

Two-sample t test (general case)

Options Help

sigma1 = 1
0 2 4 8 8 1 1.2 1.4

sigma2 = 1
0 2 4 8 8 1 1.2 1.4

Equal sigmas

n1 = 25
0 5 10 15 20 25 30 35

n2 = 25
0 5 10 15 20 25 30 35

Allocation Equal

Two-tailed Alpha .05

Equivalence

Degrees of freedom = 48

True difference of means = .5
0 .1 2 2 4 5 8 7

Power = .4101
0 2 4 8 8 1

Solve for Sample size



Power Calculator

Choose a Model and Push a Button. [Disclaimer.](#)

NORMAL	Power for a given Sample Size	Sample Size for a given Power
1 Sample	●	●
2 Sample, Equal Variances	●	●
2 Sample, Unequal Variances	●	●
Lognormal	●	●
EXPONENTIAL	Power for a given Sample Size	Sample Size for a given Power
1 Sample	●	●
2 Sample	●	●

Enter a "?" for the item to be calculated. Entering "?"s in positions 3 and 4 will calculate equal sample sizes for both groups.	
μ_1 The Mean of Population 1	<input type="text"/>
μ_2 The Mean of Population 2	<input type="text"/>
N_1 The Sample Size from Population 1	<input type="text"/>
N_2 The Sample Size from Population 2	<input type="text"/>
Sigma 1 Standard Deviation of Group 1	<input type="text"/>
Sigma 2 Standard Deviation of Group 2	<input type="text"/>
Significance Level The Significance Level of the test or Prob (reject null hypothesis ($H_0 : \mu_1 = \mu_2$) given it is true)	<input type="text"/>
Power The Power desired for the test or Prob (reject H_0 given that H_a is true)	<input type="text"/>
Number of Sides Specifies Alternative Hypothesis. One sided and $\mu_1 > \mu_2 \Rightarrow H_1 : \mu_1 > \mu_2$ One sided and $\mu_1 < \mu_2 \Rightarrow H_1 : \mu_1 < \mu_2$ Two sided $\Rightarrow H_1 : \mu_1$ not equal μ_2	<input checked="" type="radio"/> 1 Side <input type="radio"/> 2 Sides
<input type="button" value="Calculate"/>	

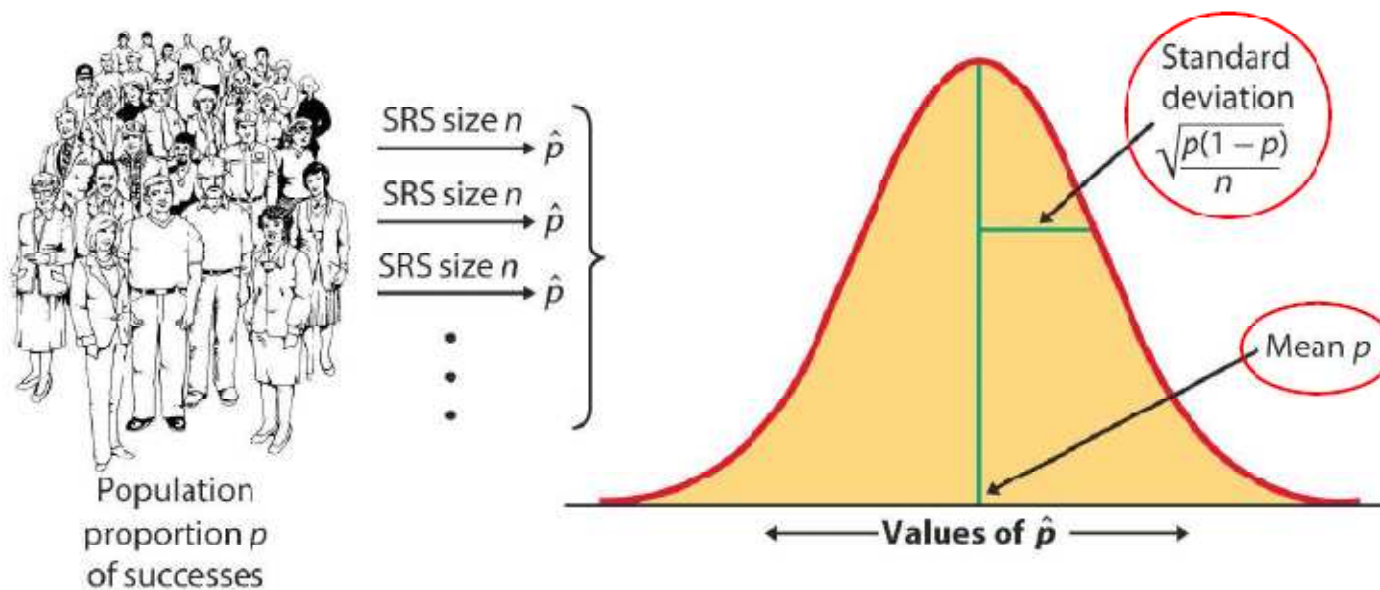
UCLA Department of Statistics

6 Inference for proportions

6.1 inference for a single proportion

Sampling distribution of a sample proportion

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.

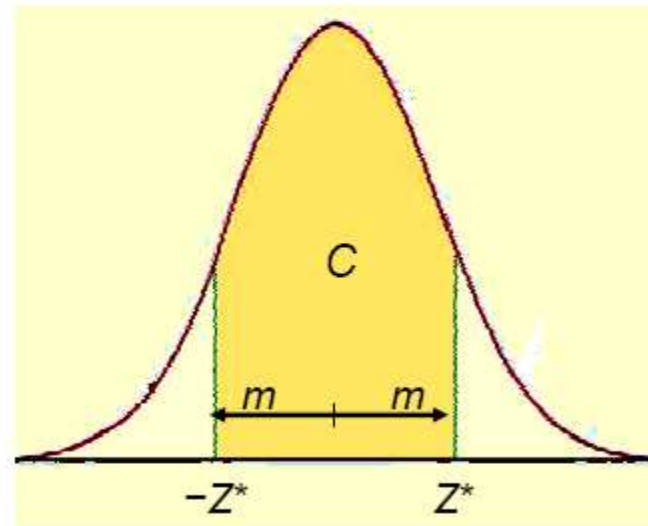
Large-sample confidence interval for p

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:

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

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

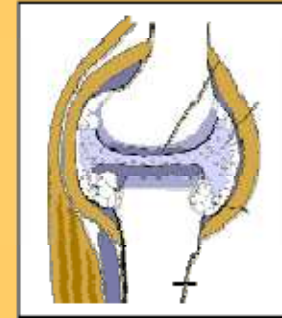
Use this method when the number of successes and the number of failures are both at least 15.



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?

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

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 :

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

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

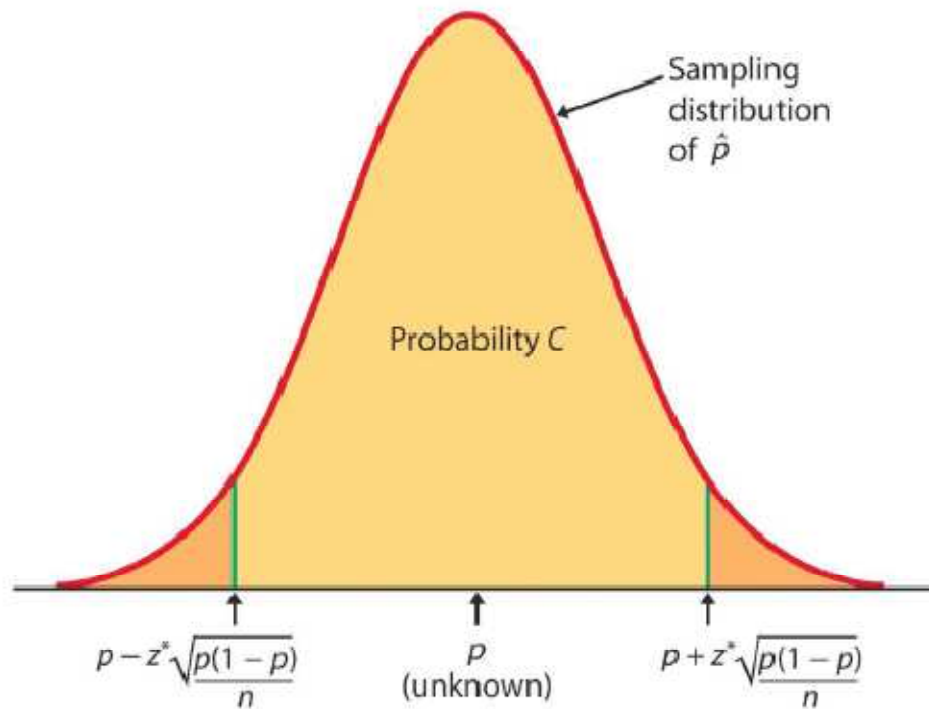
$$m = 1.645 * 0.014 \approx 0.023$$

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

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

→ With a 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!

“Plus four” confidence interval for p

A simple adjustment produces more accurate confidence intervals. We act as if we had four additional observations, two being successes and two being failures. Thus, the new sample size is $n + 4$, and the count of successes is $X + 2$.

The “plus four” estimate of p is:
$$\tilde{p} = \frac{\text{counts of successes} + 2}{\text{count of all observations} + 4}$$

And an approximate level C confidence interval is:

$CI: \tilde{p} \pm m$, with

$$m = z^* SE = z^* \sqrt{\tilde{p}(1 - \tilde{p}) / (n + 4)}$$

Use this method when C is at least 90% and sample size is at least 10.



We now use the “plus four” method to calculate the 90% confidence interval for the population proportion of arthritis patients who suffer some “adverse symptoms.”

What is the value of the “plus four” estimate of p ? $\tilde{p} = \frac{23+2}{440+4} = \frac{25}{444} \approx 0.056$

An approximate 90% confidence interval for p using the “plus four” method is:

$$m = z^* \sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$$

$$m = 1.645 * \sqrt{0.056(1-0.056) / 444}$$

$$m = 1.645 * 0.011 \approx 0.018$$

90% CI for p : $\tilde{p} \pm m$
 or 0.056 ± 0.018

→ With 90% confidence level, between 3.8% and 7.4% of arthritis patients taking this pain medication experience some adverse symptoms.

	Upper tail probability P											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.003	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
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Significance test for p

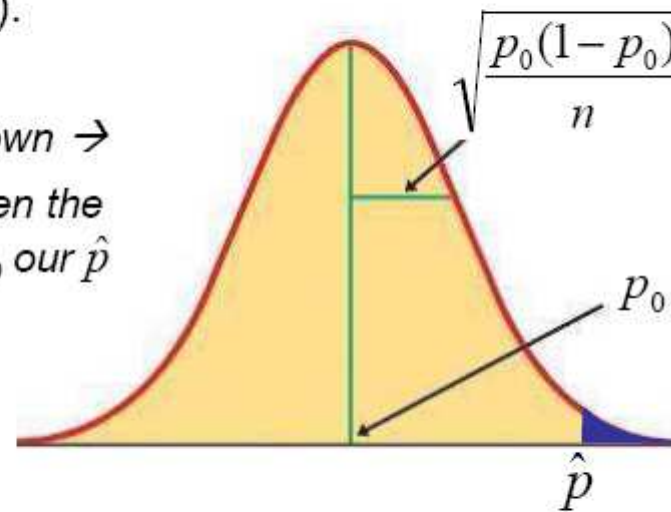
The sampling distribution for \hat{p} is approximately normal for large sample sizes and its shape depends solely on p and n .

Thus, we can easily test the null hypothesis:

$H_0: p = p_0$ (a given value we are testing).

*If H_0 is true, the sampling distribution is known \rightarrow
The likelihood of our sample proportion given the null hypothesis depends on how far from p_0 our \hat{p} is in units of standard deviation.*

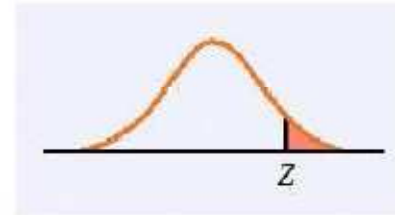
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$



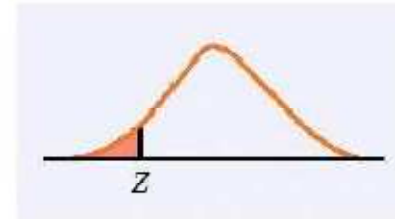
This is valid when both expected counts—expected successes np_0 and expected failures $n(1 - p_0)$ —are each 10 or larger.

P-values and one or two sided hypotheses—reminder

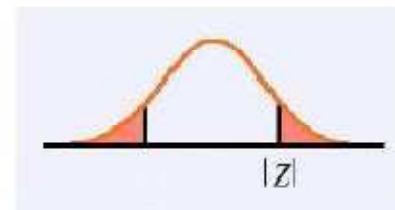
$$H_a: p > p_0 \text{ is } P(Z \geq z)$$



$$H_a: p < p_0 \text{ is } P(Z \leq z)$$



$$H_a: p \neq p_0 \text{ is } 2P(Z \geq |z|)$$



And as always, if the p-value is as small or smaller than the significance level α , then the difference is statistically significant and we reject H_0 .

A national survey by the National Institute for Occupational Safety and Health on restaurant employees found that 75% said that work stress had a negative impact on their personal lives.

You investigate a restaurant chain to see if the proportion of all their employees negatively affected by work stress differs from the national proportion $p_0 = 0.75$.

$$H_0: p = p_0 = 0.75 \text{ vs. } H_a: p \neq 0.75 \text{ (2 sided alternative)}$$

In your SRS of 100 employees, you find that 68 answered “Yes” when asked, “Does work stress have a negative impact on your personal life?”

The expected counts are $100 \times 0.75 = 75$ and 25.

Both are greater than 10, so we can use the z-test.

The test statistic is:

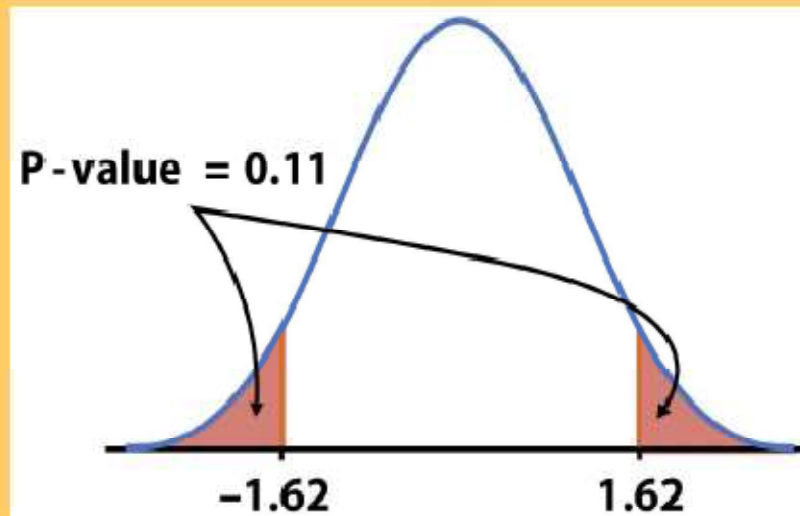
$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \\ &= \frac{0.68 - 0.75}{\sqrt{\frac{(0.75)(0.25)}{100}}} = 1.62 \end{aligned}$$

From Table A we find the area to the left of $z = 1.62$ is 0.9474.

Thus $P(Z \geq 1.62) = 1 - 0.9474$, or 0.0526. Since the alternative hypothesis is two-sided, the P -value is the area in both tails, and $P = 2 \times 0.0526 = 0.1052$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

→ The chain restaurant data are not significantly different from the national survey results ($\hat{p} = 0.68$, $z = 1.62$, $P = 0.11$).



Software gives you summary data (sample size and proportion) as well as the actual p-value.

Minitab

Test and Confidence Interval for One Proportion

Test of $p = 0.75$ vs $p \text{ not} = 0.75$

Sample	X	N	Sample p	95.0 % CI	Z-Value	P-Value
1	68	100	0.680000	(0.588572, 0.771428)	-1.62	0.106

Crunch It!

Hypothesis test results:

p = proportion of successes for population

Parameter: p

H0 : Parameter = 0.75

HA : Parameter not = 0.75

Proportion	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
p	68	100	0.68	0.04330127	-1.6165807	0.106

Interpretation: magnitude versus 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^* .

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

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							

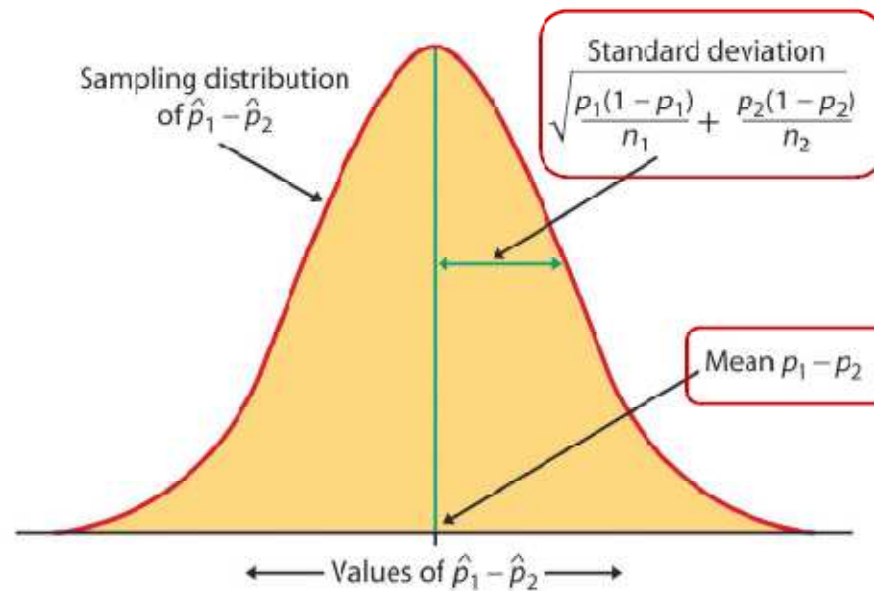
$$n = \left(\frac{z^*}{m} \right)^2 p^* (1 - p^*) = \left(\frac{1.645}{0.01} \right)^2 (0.1)(0.9) \approx 2434.4$$

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

6.2 Comparing two proportions

Comparing two independent samples

We often need to compare two treatments used on **independent** samples. We can compute the difference between the two sample proportions and compare it to the corresponding, approximately normal sampling distribution for $(\hat{p}_1 - \hat{p}_2)$:



Large-sample confidence interval for two proportions

For two independent SRSs of sizes n_1 and n_2 with sample proportion of successes \hat{p}_1 and \hat{p}_2 respectively, an **approximate level C confidence interval** for $p_1 - p_2$ is

$(\hat{p}_1 - \hat{p}_2) \pm m$, m is the margin of error

$$m = z^* SE_{diff} = z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

C is the area under the standard normal curve between $-z^*$ and z^* .

Use this method only when the populations are at least 10 times larger than the samples and the number of successes and the number of failures are each at least 10 in each samples.

Cholesterol and heart attacks

How much does the cholesterol-lowering drug Gemfibrozil help reduce the risk of heart attack? We compare the incidence of heart attack over a 5-year period for two random samples of middle-aged men taking either the drug or a placebo.

Standard error of the difference $p_1 - p_2$:

$$SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$SE = \sqrt{\frac{0.0273(0.9727)}{2051} + \frac{0.0414(0.9586)}{2030}} = 0.00764$$

	H. attack	n	\hat{p}
Drug	56	2051	2.73%
Placebo	84	2030	4.14%

The confidence interval is $(\hat{p}_1 - \hat{p}_2) \pm z * SE$

So the 90% CI is $(0.0414 - 0.0273) \pm 1.645 * 0.00746 = 0.0141 \pm 0.0125$

We are 90% confident that the percentage of middle-aged men who suffer a heart attack is 0.16% to 2.7% lower when taking the cholesterol-lowering drug.

“Plus four” confidence interval for two proportions

The “plus four” method again produces more accurate confidence intervals. We act as if we had four additional observations: one success and one failure in each of the two samples. The new combined sample size is $n_1 + n_2 + 4$ and the proportions of successes are:

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

An approximate level C confidence interval is:

$$CI: (\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

Use this when C is at least 90% and both sample sizes are at least 5.

Cholesterol and heart attacks

Let's now calculate the "plus four" CI for the difference in percentage of middle-aged men who suffer a heart attack (placebo – drug).

	H. attack	n	\tilde{p}
Drug	56	2051	2.78%
Placebo	84	2030	4.18%

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} = \frac{56 + 1}{2051 + 2} \approx 0.0278 \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2} = \frac{84 + 1}{2030 + 2} \approx 0.0418$$

Standard error of the population difference $p_1 - p_2$:

$$SE = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}} = \sqrt{\frac{0.0278(0.9722)}{2053} + \frac{0.0418(0.9582)}{2032}} = 0.0057$$

The confidence interval is $(\tilde{p}_1 - \tilde{p}_2) \pm z * SE$

So the 90% CI is $(0.0418 - 0.0278) \pm 1.645 * 0.00573 = 0.014 \pm 0.0094$

We are 90% confident that the percentage of middle-aged men who suffer a heart attack is 0.46% to 2.34% lower when taking the cholesterol-lowering drug.

Test of significance

If the null hypothesis is true, then we can rely on the properties of the sampling distribution to estimate the probability of drawing 2 samples with proportions \hat{p}_1 and \hat{p}_2 at random.

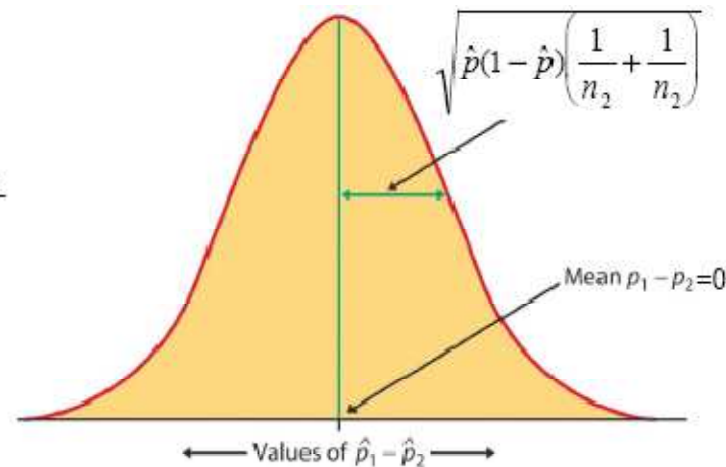
$$H_0 : p_1 = p_2 = p$$

Our best estimate of p is \hat{p} ,

the pooled sample proportion

$$\hat{p} = \frac{\text{total successes}}{\text{total observations}} = \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



This test is appropriate when the populations are at least 10 times as large as the samples and all counts are at least 5 (number of successes and number of failures in each sample).

Gastric Freezing

Gastric freezing was once a treatment for ulcers. Patients would swallow a deflated balloon with tubes, and a cold liquid would be pumped for an hour to cool the stomach and reduce acid production, thus relieving ulcer pain. **The treatment was shown to be safe, significantly reducing ulcer pain** and widely used for years.



A randomized comparative experiment later compared the outcome of gastric freezing with that of a placebo: 28 of the 82 patients subjected to gastric freezing improved, while 30 of the 78 in the control group improved.

$$H_0: p_{gf} = p_{\text{placebo}} \quad \hat{p}_{\text{pooled}} = \frac{28+30}{82+78} = 0.3625$$

$$H_a: p_{gf} > p_{\text{placebo}}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.341 - 0.385}{\sqrt{0.363 * 0.637\left(\frac{1}{82} + \frac{1}{78}\right)}} = \frac{-0.044}{\sqrt{0.231 * 0.025}} = -0.499$$

Conclusion: The gastric freezing was no better than a placebo (p-value 0.69), and this treatment was abandoned. **ALWAYS USE A CONTROL!**

Relative risk

Another way to compare two proportions is to study the ratio of the two proportions, which is often called the **relative risk (RR)**. A relative risk of 1 means that the two proportions are equal.

The procedure for calculating confidence intervals for relative risk is more complicated (use software) but still based on the same principles that we have studied.

The age at which a woman gets her first child may be an important factor in the risk of later developing breast cancer. An international study selected women with at least one birth and recorded if they had breast cancer or not and whether they had their first child before their 30th birthday or after.

	Birth age 30+	Sample size	\hat{p}
Cancer	683	3220	21.2%
No	1498	10,245	14.6%

$$RR = \frac{.212}{.146} \approx 1.45$$

Women with a late first child have 1.45 times the risk of developing breast cancer.

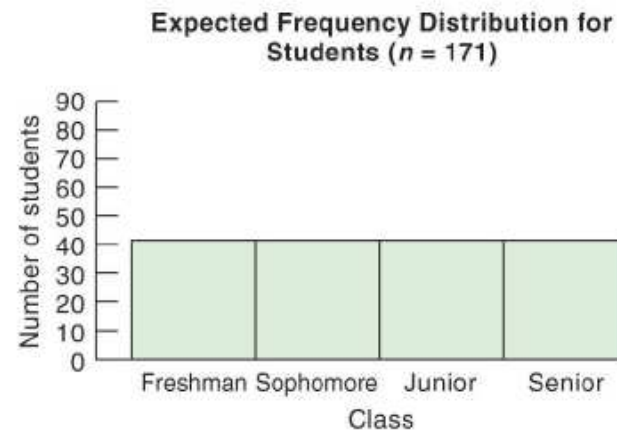
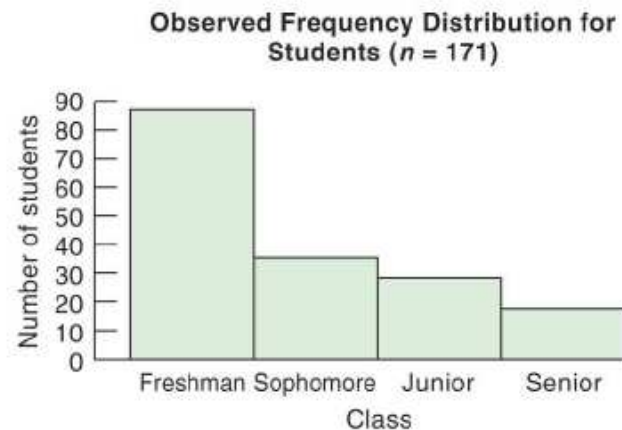
7 Analysis of two-way tables

7.1 Inference for two-way tables

Hypothesis: no association

Again, we want to know if the differences in sample proportions are likely to have occurred just by chance due to random sampling.

We use the **chi-square (χ^2) test** to assess the null hypothesis of no relationship between the two categorical variables of a two-way table.



Expected cell count

Two-way tables sort the data according to two categorical variables. We want to test the hypothesis that there is no relationship between these two categorical variables (H_0).

To test this hypothesis, we compare **actual counts** from the sample data with **expected counts**, given the null hypothesis of no relationship.

The expected count in any cell of a two-way table when H_0 is true is:

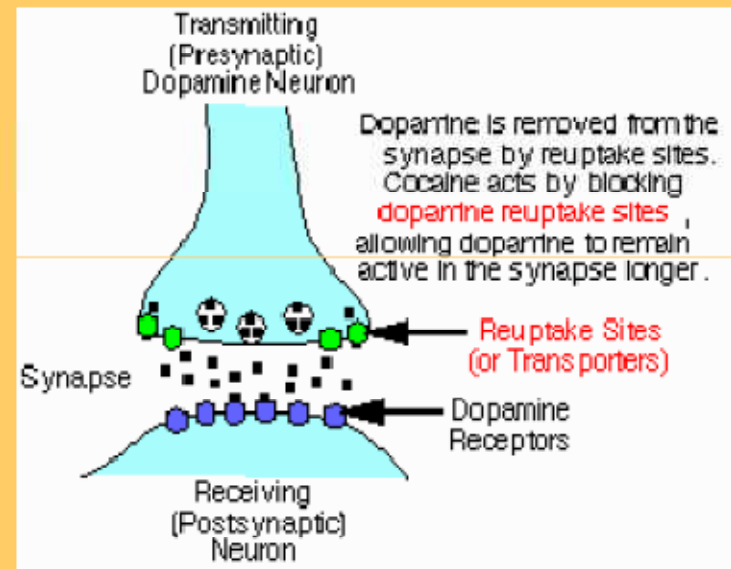
$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

Cocaine addiction

Cocaine produces short-term feelings of physical and mental well being. To maintain the effect, the drug may have to be taken more frequently and at higher doses. After stopping use, users will feel tired, sleepy and **depressed**.

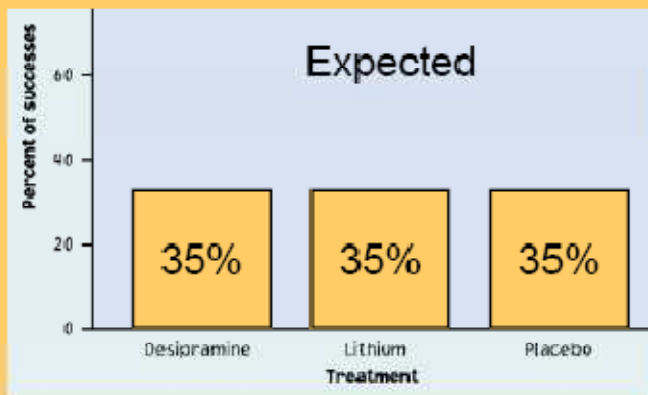
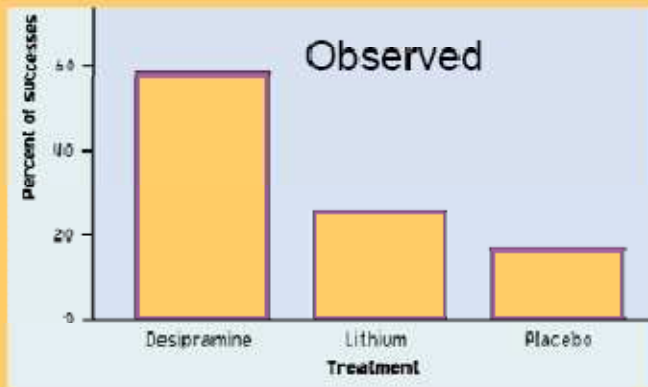
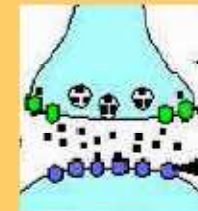
The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

Desipramine is an antidepressant affecting the brain chemicals that may become unbalanced and cause depression. It was thus tested for recovery from cocaine addiction.



Treatment with desipramine was compared to a standard treatment (lithium, with strong anti-manic effects) and a placebo.

Cocaine addiction



	Relapse		Total
	No	Yes	
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74

Expected relapse counts

	No	Yes
Desipramine	$25 \cdot \frac{26}{74} \approx 8.78$ $25 \cdot 0.35$	16.22 $25 \cdot 0.65$
Lithium	9.14 $26 \cdot 0.35$	16.86 $26 \cdot 0.65$
Placebo	8.08 $23 \cdot 0.35$	14.92 $23 \cdot 0.65$

The chi-square test

The chi-square statistic (χ^2) is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts.

The formula for the χ^2 statistic is:
(summed over all $r \times c$ cells in the table)

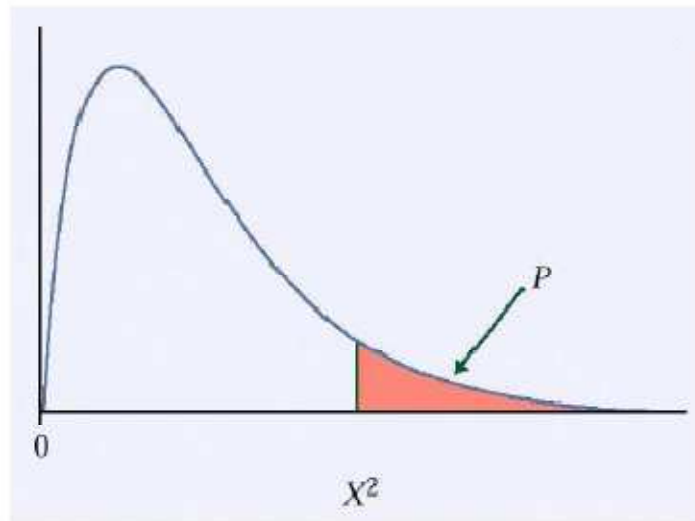
$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Large values for χ^2 represent strong deviations from the expected distribution under the H_0 and provide evidence against H_0 .

However, since χ^2 is a sum, how large a χ^2 is required for statistical significance will depend on the number of comparisons made.

For the chi-square test, H_0 states that there is no association between the row and column variables in a two-way table. The alternative is that these variables are related.

If H_0 is true, the chi-square test has approximately a χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom.



The P-value for the chi-square test is the area to the right of χ^2 under the χ^2 distribution with df $(r-1)(c-1)$:

$$P(\chi^2 \geq X^2).$$

When is it safe to use a chi-square test?

We can safely use the chi-square test when:

- The samples are simple random samples (**SRS**).
- All individual **expected counts** are 1 or more (≥ 1)
- No more than 20% of **expected counts** are less than 5 (< 5)
 - ➔ *For a 2x2 table, this implies that all four expected counts should be 5 or more.*

Chi-square test versus z-test for two proportions

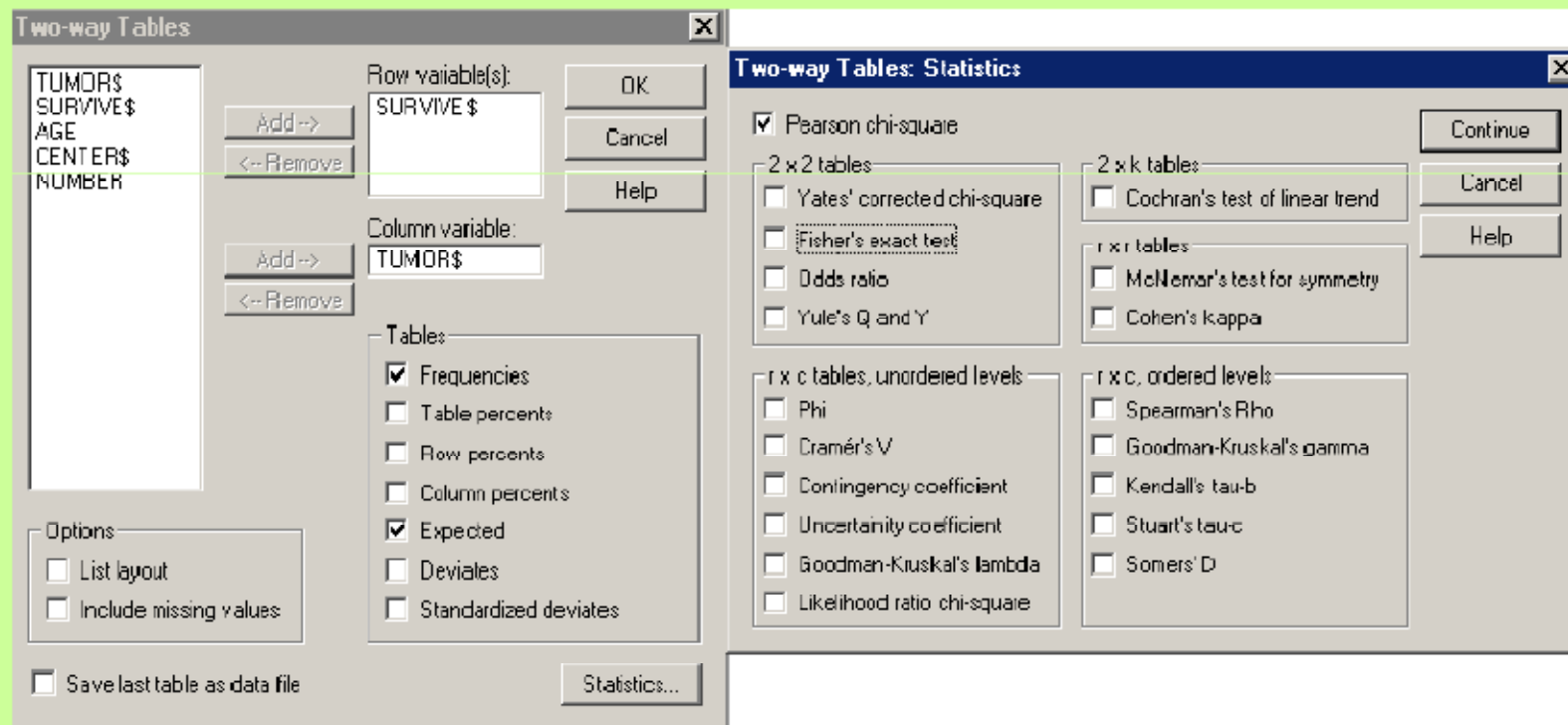
When comparing only two proportions, such as in a 2x2 table where the columns represent counts of “success” and “failure,” we can test

$$H_0: p_1 = p_2 \text{ vs. } H_a: p_1 \neq p_2$$

equally with a two-sided z test or with a chi-square test with 1 degree of freedom and get the same p-value. In fact, the two test statistics are related: $\chi^2 = (z)^2$.

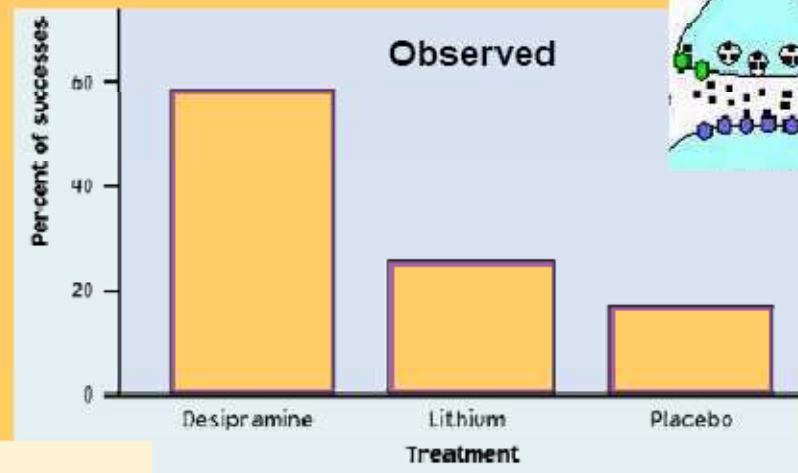
Using software

- ❑ In Excel you have to do almost all the calculations for the chi-square test yourself, and it only gives you the p-value (not the component).
- ❑ This is **Systat**: Menu/Statistics/Crosstabs



Cocaine addiction

Minitab statistical software output for the cocaine study



Chi-Square Test

Expected counts are printed below observed counts

	Success	Relapse	Total
D	14	10	24
	8.00	16.00	
L	6	18	24
	8.00	16.00	
P	4	20	24
	8.00	16.00	
Total	24	48	72
Chi-Sq	= 4.500 + 2.250 + 0.500 + 0.250 + 2.000 + 1.000	= 10.500	
DF	= 2, P-Value = 0.005		

The p-value is 0.005 or half a percent. This is very significant.

We reject the null hypothesis of no association and conclude that there is a significant relationship between treatment (*desipramine, lithium, placebo*) and outcome (*relapse or not*).

Successful firms

Franchise businesses are sometimes given an exclusive territory by contract. This means that the new outlet will not have to compete with other outlets of the same chain within its own territory. How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

A random sample of 170 new franchises recorded two categorical variables for each firm: (1) whether the firm was successful or not (based on economic criteria) and (2) whether or not the firm had an exclusive-territory contract.

Observed numbers of firms			
	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

This is a 2x2 table (two levels for success, yes/no; two levels for exclusive territory, yes/no).

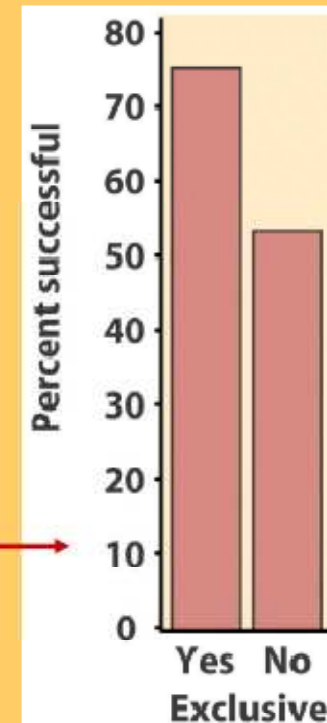
$$\rightarrow df = (2 - 1)(2 - 1) = 1$$

Successful firms

How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

To compare firms that have an exclusive territory with those that do not, we start by examining column percents (conditional distribution):

Column percents for firms		
	Exclusive territory	
Success	Yes	No
Yes	76%	54%
No	24%	46%
Total	100%	100%



The difference between the percent of successes among the two types of firms is quite large. The chi-square test can tell us whether or not these differences can be plausibly attributed to chance (random sampling). Specifically, we will test

H_0 : No relationship between exclusive clause and success

H_a : There is some relationship between the two variables

Successful firms

Here is the chi-square output from **Minitab**:

```

Rows: Success      Columns: Excl
      1_Yes      2_No      All
1_Yes   108      15      123
      102.74    20.26    123.00
2_No    34       13      47
      39.26     7.74     47.00
All     142      28     170
      142.00    28.00    170.00

Chi - Square = 5.911, DF = 1, P -Value = 0.015

Cell Contents  --
               Count
               Exp Freq

```

The p-value is significant at $\alpha = 5\%$ ($p = 1.5\%$) thus we reject H_0 : we have found a significant relationship between an exclusive territory and the success of a franchised firm.

Successful firms

	Yes	No	Total
Yes	108 87.8% 76.06% 63.53%	15 12.2% 53.57% 8.824%	123 100.00% 72.35% 72.35%
No	34 72.34% 23.94% 20%	13 27.66% 46.43% 7.647%	47 100.00% 27.65% 27.65%
Total	142 83.53% 100.00% 83.53%	28 16.47% 100.00% 16.47%	170 100.00% 100.00% 100.00%

Computer output
using **Crunch It!**

Cell format:

Count
Row percent
Column percent
Total percent

Test for independence of Success and Exclusive Territory:

Statistic	DF	Value	P-value
Chi-square	1	5.9111857	0.015

7.2 Formulas and models for two-way tables and Goodness-of-fit

Computations for two-way tables

When analyzing relationships between two categorical variables, follow this procedure:

1. Calculate descriptive statistics that convey the important information in the table—usually column or row percents.
2. Find the expected counts and use them to compute the X^2 statistic.
3. Compare your X^2 statistic to the chi-square critical values from Table F to find the approximate P -value for your test.
4. Draw a conclusion about the association between the row and column variables.

Computing conditional distribution

The calculated percents within a two-way table represent the **conditional distributions** describing the “relationship” between both variables.

For every two-way table, there are two sets of possible conditional distributions (column percents or row percents).

For column percents, divide each cell count by the column total. The sum of the percents in each column should be 100, except for possible small roundoff errors.

When one variable is clearly explanatory, it makes sense to describe the relationship by comparing the conditional distributions of the response variable for each value (level) of the explanatory variable.

Music and wine purchase decision

What is the relationship between type of music played in supermarkets and type of wine purchased?

We want to compare the conditional distributions of the response variable (wine purchased) for each value of the explanatory variable (music played). Therefore, we calculate column percents.

Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine. $30/84 = 0.357 \rightarrow 35.7\%$ of the wine sold was French when no music was played.



We calculate the column conditional percents similarly for each of the nine cells in the table:

Wine	Music			Total
	None	French	Italian	
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

$$\frac{30}{84} = 35.7\%$$

$$= \frac{\text{cell total}}{\text{column total}}$$

Column percents for wine and music

Wine	Music			Total
	None	French	Italian	
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

Computing expected counts

When testing the null hypothesis that there is no relationship between both categorical variables of a two-way table, we compare **actual counts** from the sample data with **expected counts** given H_0 .

The expected count in any cell of a two-way table when H_0 is true is:

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

Although in real life counts must be whole numbers, an expected count need not be. The expected count is the mean over many repetitions of the study, assuming no relationship.

Music and wine purchase decision

The null hypothesis is that there is no relationship between music and wine sales. The alternative is that these two variables are related.

Wine	Music			Total
	None	French	Italian	
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

What is the expected count in the upper-left cell of the two-way table, under H_0 ?

Column total 84: Number of bottles sold without music

Row total 99: Number of bottles of French wine sold

Table total 243: all bottles sold during the study

This expected cell count is thus $(84)(99) / 243 = 34.222$



Nine similar calculations produce the table of expected counts:

Wine	Music			Total
	None	French	Italian	
French	34.222	30.556	34.222	99.000
Italian	10.716	9.568	10.716	31.000
Other	39.062	34.877	39.062	113.001
Total	84.000	75.001	84.000	243.001

Computing the chi-square statistic

The chi-square statistic (χ^2) is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts.

The formula for the χ^2 statistic is:

(summed over all $r \times c$ cells in the table)

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Tip: First, calculate the χ^2 components, $(\text{observed} - \text{expected})^2 / \text{expected}$, for each cell of the table, and then sum them up to arrive at the χ^2 statistic.

Music and wine purchase decision

H_0 : No relationship between music and wine

H_a : Music and wine are related

Observed counts

Wine	Music		
	None	French	Italian
French	30	39	30
Italian	11	1	19
Other	43	35	35

Expected counts

Wine	Music		
	None	French	Italian
French	34.222	30.556	34.222
Italian	10.716	9.568	10.716
Other	39.062	34.877	39.062

We calculate nine χ^2 components and sum them to produce the χ^2 statistic:



$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(30 - 34.222)^2}{34.222} + \frac{(39 - 30.556)^2}{30.556} + \frac{(30 - 34.222)^2}{34.222} \\
 &\quad + \frac{(11 - 10.716)^2}{10.716} + \frac{(1 - 9.568)^2}{9.568} + \frac{(19 - 10.716)^2}{10.716} \\
 &\quad + \frac{(43 - 39.062)^2}{39.062} + \frac{(35 - 34.877)^2}{34.877} + \frac{(35 - 39.062)^2}{39.062} \\
 &= 0.5209 + 2.3337 + 0.5209 + 0.0075 + 7.6724 \\
 &\quad + 6.4038 + 0.3971 + 0.0004 + 0.4223 \\
 &= 18.28
 \end{aligned}$$

Finding the p-value with Table F

The χ^2 distributions are a family of distributions that can take only positive values, are skewed to the right, and are described by a specific degrees of freedom.

Table F gives upper critical values for many χ^2 distributions.

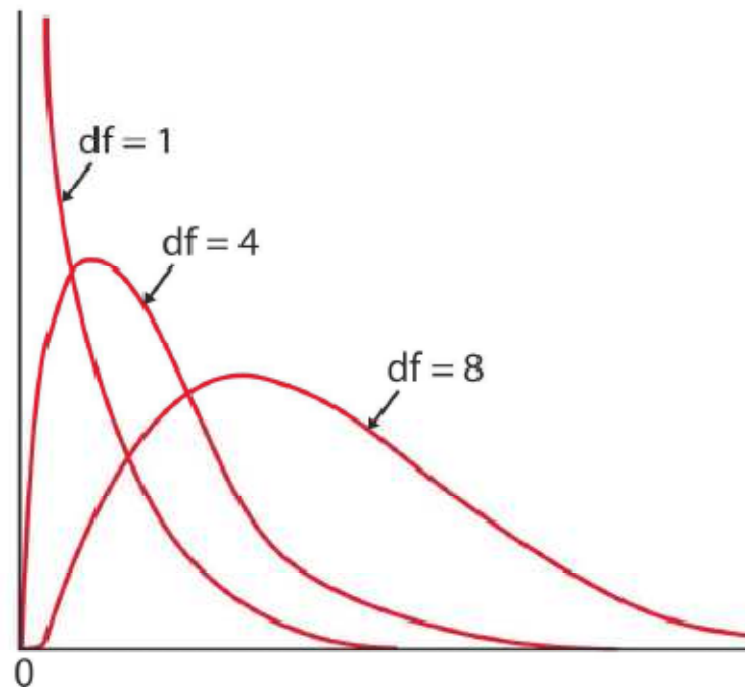


Table F



$df = (r-1)(c-1)$

Ex: In a 4x3 table,
 $df = 3*2 = 6$

If $\chi^2 = 16.1$,
the p-value
is between
0.01-0.02.

df	p												
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005	
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12	
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20	
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73	
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00	
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11	
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10	
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02	
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87	
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67	
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42	
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14	
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82	
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48	
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11	
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72	
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31	
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88	
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43	
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97	
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50	
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01	
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51	
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00	
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48	
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95	
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41	
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86	
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30	
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73	
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16	
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09	
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56	
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.70	
80	88.13	90.41	93.11	96.58	101.90	106.60	108.10	112.30	116.30	120.10	124.80	128.30	
100	109.10	111.70	114.70	118.50	124.30	129.60	131.10	135.80	140.20	144.30	149.40	153.20	

Music and wine purchase decision

H_0 : No relationship between music and wine H_a : Music and wine are related

Wine	Music		
	None	French	Italian
French	30	39	30
Italian	11	1	19
Other	43	35	35

We found that the χ^2 statistic under H_0 is 18.28.

The two-way table has a 3x3 design (3 levels of music and 3 levels of wine). Thus, the degrees of freedom for the χ^2 distribution for this test is:

$$(r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$$

df	p												
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005	
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12	
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20	
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73	
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00	

$$16.42 < \chi^2 = 18.28 < 18.47$$

$$0.0025 > \text{p-value} > 0.001 \rightarrow \text{very significant}$$

There is a significant relationship between the type of music played and wine purchases in supermarkets.



Interpreting the chi-square output

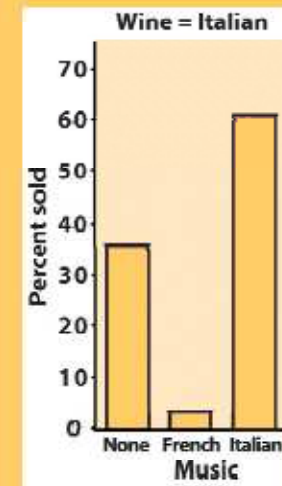
- The values summed to make up χ^2 are called the **χ^2 components**.
When the test is statistically significant, the **largest components** point to the conditions most different from the expectations based on H_0 .

Music and wine purchase decision

χ^2 components	Music		
	None	French	Italian
Wine			
French	0.5209	2.3337	0.5209
Italian	0.0075	7.6724	6.4038
Other	0.3971	0.0004	0.4223



Two chi-square components contribute most to the χ^2 total → the largest effect is for sales of Italian wine, which are strongly affected by Italian and French music.



Actual proportions show that Italian music helps sales of Italian wine, but French music hinders it.

Models for two-way tables

The chi-square test is an overall technique for comparing any number of population proportions, testing for evidence of a relationship between two categorical variables. We can either:

- ▣ **Compare several populations:** Randomly select several SRSs each from a different population (or from a population subjected to different treatments) → experimental study.
- ▣ **Test for independence:** Take one SRS and classify the individuals in the sample according to two categorical variables (attribute or condition) → observational study, historical design.

Both models use the χ^2 test to test the hypothesis of *no relationship*.

Comparing several populations

Select independent SRSs from each of c populations, of sizes n_1, n_2, \dots, n_c . Classify each individual in a sample according to a categorical response variable with r possible values. There are c different probability distributions, one for each population.

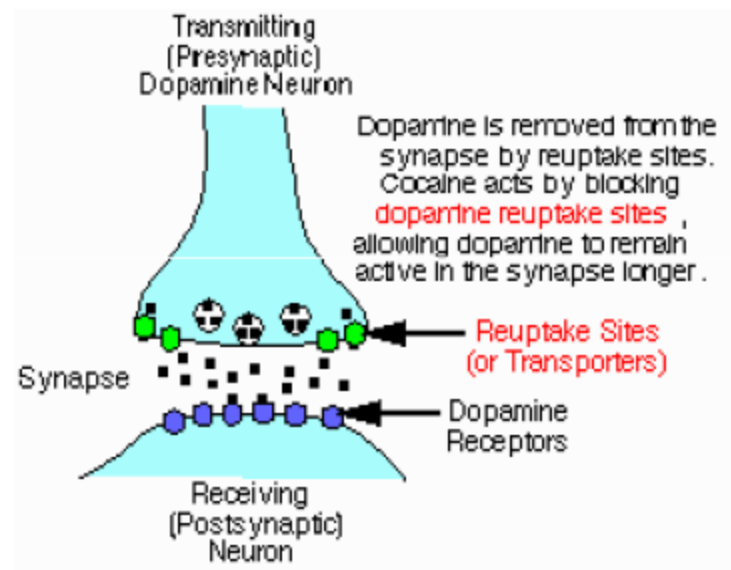
The null hypothesis is that the distributions of the response variable are the same in all c populations. The alternative hypothesis says that these c distributions are not all the same.

Cocaine addiction

Cocaine produces short-term feelings of physical and mental well being. To maintain the effect, the drug may have to be taken more frequently and at higher doses. After stopping use, users will feel tired, sleepy, and **depressed**.

The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

We compare treatment with an anti-depressant (desipramine), a standard treatment (lithium), and a placebo.



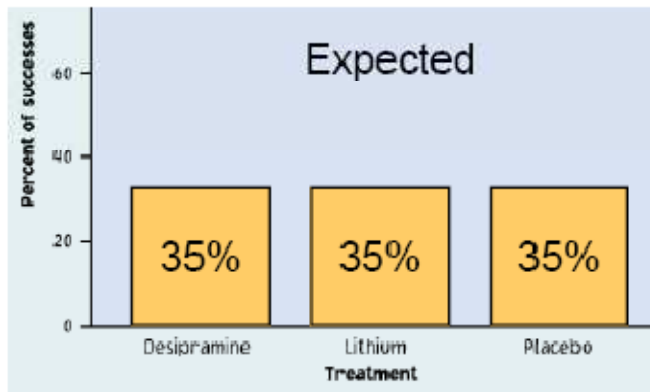
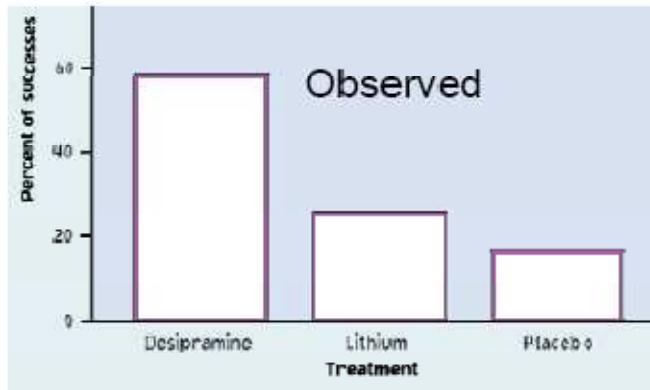
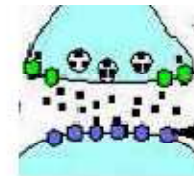
Population 1: Antidepressant treatment (desipramine)

Population 2: Standard treatment (lithium)

Population 3: Placebo ("sugar pill")

Cocaine addiction

H_0 : The proportions of success (no relapse) are the same in all three populations.



	Relapse		Total
	No	Yes	
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74

Expected relapse counts

	No	Yes
Desipramine	$25 \cdot 26 / 74 \approx 8.78$ $25 \cdot 0.35$	16.22 $25 \cdot 0.65$
Lithium	9.14 $26 \cdot 0.35$	16.86 $25 \cdot 0.65$
Placebo	8.08 $23 \cdot 0.35$	14.92 $25 \cdot 0.65$

Cocaine addiction

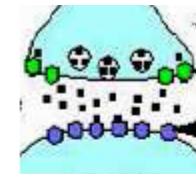


Table of counts:
 “actual / **expected**,” with
 three rows and two
 columns:

$$df = (3-1)*(2-1) = 2$$

	No relapse	Relapse
Desipramine	15 8.78	10 16.22
Lithium	7 9.14	19 16.86
Placebo	4 8.08	19 14.92

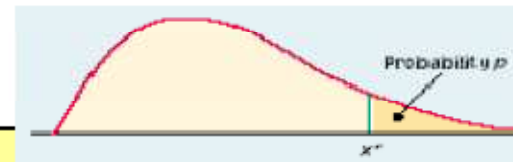
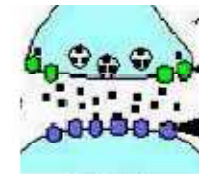
$$\begin{aligned} \chi^2 &= \frac{(15 - 8.78)^2}{8.78} + \frac{(10 - 16.22)^2}{16.22} \\ &+ \frac{(7 - 9.14)^2}{9.14} + \frac{(19 - 16.86)^2}{16.86} \\ &+ \frac{(4 - 8.08)^2}{8.08} + \frac{(19 - 14.92)^2}{14.92} \\ &= 10.74 \end{aligned}$$

χ^2 components:

$$\begin{array}{r} \longrightarrow \\ \begin{array}{cc} 4.41 & 2.39 \\ 0.50 & 0.27 \\ 2.06 & 1.12 \end{array} \end{array}$$

Cocaine addiction: Table F

H_0 : The proportions of success (no relapse) are the same in all three populations.



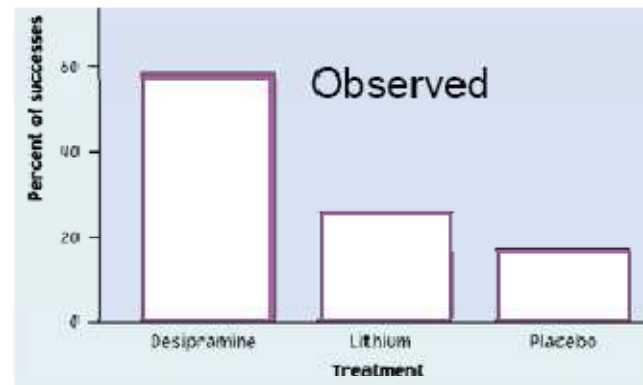
df	p											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60★	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

$\chi^2 = 10.71$ and $df = 2$

$10.60 < \chi^2 < 11.98 \rightarrow 0.005 < p < 0.0025 \rightarrow$ reject the H_0

→ The proportions of success are not the same in all three populations (Desipramine, Lithium, Placebo).

Desipramine is a more successful treatment →→→



Testing for independence

We now have a *single* sample from a *single* population. For each individual in this SRS of size n we measure two categorical variables. The results are then summarized in a two-way table.

The null hypothesis is that the row and column variables are independent. The alternative hypothesis is that the row and column variables are dependent.

Successful firms

How does the presence of an exclusive-territory clause in the contract for a franchise business relate to the survival of that business?

A random sample of 170 new franchises recorded two categorical variables for each firm: (1) whether the firm was successful or not (based on economic criteria) and (2) whether or not the firm had an exclusive-territory contract.

Observed numbers of firms			
	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

This is a 2x2, two-way table

(2 levels for business success, yes/no,
2 levels for exclusive territory, yes/no).

We will test H_0 : The variables exclusive clause and success are independent.

Successful firms

Computer output for
the chi-square test
using **Minitab**:

```

Rows : Success      Columns : Excl
      1_Yes         2_No         All
1_Yes  108          15           123
      102.74        20.26        123.00
2_No   34           13           47
      39.26         7.74         47.00
All    142          28           170
      142.00        28.00        170.00

Chi-Square = 5.911, DF = 1, P -Value = 0.015

Cell Contents  --
                Count
                Exp Freq

```

The p-value is significant at α 5% thus we reject H_0 :

The existence of an exclusive territory clause in a franchise's contract and the success of that franchise are not independent variables.

Parental smoking

Does parental smoking influence the incidence of smoking in children when they reach high school? Randomly chosen high school students were asked whether they smoked (columns) and whether their parents smoked (rows).

Examine the computer output for the chi-square test performed on these data. What does it tell you?

Sample size?

Hypotheses?

Are data ok for χ^2 test?

Interpretation?

Chi-Square Test			
Expected counts are printed below observed counts			
	Smokes	NoSmoke	Total
Both	400 332.49	1380 1447.51	1780
One	416 418.22	1823 1820.78	2239
None	188 253.29	1168 1102.71	1356
Total	1004	4371	5375
Chi-Sq =	13.709	+ 3.149	+ 0.012
		+ 0.003	+ 16.829
		+ 3.866	= 37.566
DF = 2,	P-Value = 0.000		

Testing for Goodness-of-fit

We have used the chi-square test as the tool to compare two or more distributions all based on sample data.

We now consider a slight variation on this scenario where only one of the distributions is known (our sample data observations) and we want to compare it with a hypothesized distribution.

- ▣ Data for n observations on a categorical variable with k possible outcomes are summarized as observed counts, n_1, n_2, \dots, n_k .
- ▣ The null hypothesis specifies probabilities p_1, p_2, \dots, p_k for the possible outcomes.

Car accidents and day of the week

A study of 667 drivers who were using a cell phone when they were involved in a collision on a weekday examined the relationship between these accidents and the day of the week.

Number of collisions by day of the week					
Day of the week					
Mon.	Tue.	Wed.	Thu.	Fri.	Total
133	126	159	136	113	667

Are the accidents equally likely to occur on any day of the working week?

H_0 specifies that all 5 days are equally likely for car accidents \rightarrow each $p_i = 1/5$.

The chi-square goodness-of-fit test

Data for n observations on a categorical variable with k possible outcomes are summarized as observed counts, n_1, n_2, \dots, n_k in k cells.

H_0 specifies probabilities p_1, p_2, \dots, p_k for the possible outcomes.

For each cell, multiply the total number of observations n by the specified probability p_i :

$$\text{expected count} = np_i$$

The **chi-square statistic** follows the chi-square distribution with **$k - 1$ degrees of**

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Car accidents and day of the week

H_0 specifies that all days are equally likely for car accidents → each $p_i = 1/5$.

Number of collisions by day of the week

Day of the week					
Mon.	Tue.	Wed.	Thu.	Fri.	Total
133	126	159	136	113	667

The expected count for each of the five days is $np_i = 667(1/5) = 133.4$.

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(\text{count}_{\text{day}} - 133.4)^2}{133.4} = 8.49$$

Following the chi-square distribution with $5 - 1 = 4$ degrees of freedom.

df	p											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

The p-value is thus between 0.1 and 0.05, which is not significant at $\alpha 5\%$.

→ There is no significant evidence of different car accident rates for different weekdays when the driver was using a cell phone.

Cautionary note about software

The chi-square function in **Excel** does not compute expected counts automatically but instead lets you provide them. This makes it easy to test for goodness of fit. You then get the test's p-value—but no details of the X^2 calculations.

=CHITEST(array of actual values, array of expected values)
with values arranged in two similar $r \times c$ tables
--> returns the p value of the Chi Square test

Note: Many software packages do not provide a direct way to compute the chi-square goodness of fit test. But you can find a way around:

Make a two-way table where the first column contains k cells with the observed counts. Make a second column with counts that correspond *exactly* to the probabilities specified by H_0 , using a very large number of observations. Then analyze the two-way table with the chi-square function.

8 How to select the appropriate test?

(<http://www.graphpad.com/www/book/choose.htm>)

	Type of Data			
Goal	Measurement (from Gaussian Population)	Rank, Score, or Measurement (from Non- Gaussian Population)	Binomial (Two Possible Outcomes)	Survival Time
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*

Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression

References:

- Mood
- Freeman IPS slides 4