Hypothesis testing

Natalie Hunt

galvanize

Objectives

- State the difference between hypothesis testing and parameter estimation
- State the steps of hypothesis testing
- Know when to use a Z or t-test
- State type I and type II error in your own words
- Be able to compare two sample means and two sample proportions
- State when to use the Bonferroni correction

Hypothesis Testing

Hypothesis testing allows us to systematically quantify how certain we can be in interpreting the results of a statistical experiment.

An experiment is any situation where you take a random sample of a population and measure something about it.

What are some examples of hypothesis tests?

5 Steps of Hypothesis Testing

1. State the **null hypothesis** (H_0) and **alternative hypothesis** (H_A)
2. Choose a **significance level** (α), often .05
3. Select statistical test, and compute appropriate **test statistic**
4. Compute **p-value** based on test statistic
5. State **conclusion**
    - If p-value < α: Reject H_0
    - If p-value >= α: Fail to reject H_0

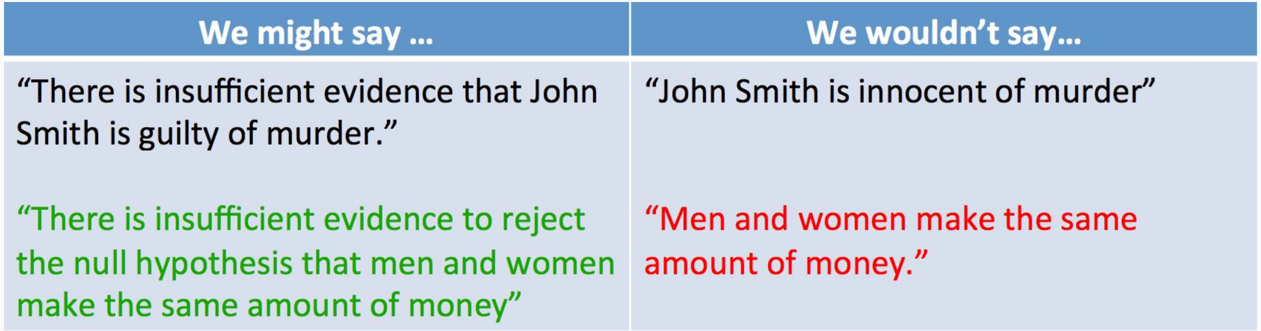
Step 1: Null and Alternative Hypotheses

The set-up: **Innocent (H_0) until proven guilty (H_A)**

In other words, state the most conservative position as the null hypothesis. Scientists are skeptics first!

Or think of the alternative hypothesis as the question you are seeking to confirm/disconfirm. For example:

- H_0: Men and women make the same amount of money
- H_A: Men and women do not make the same amount of money



We might say …	We wouldn’t say…
“There is insufficient evidence that John Smith is guilty of murder.” “There is insufficient evidence to reject the null hypothesis that men and women make the same amount of money”	“John Smith is innocent of murder” “Men and women make the same amount of money.”

Step 2: Choose a Significance Level

Error types:

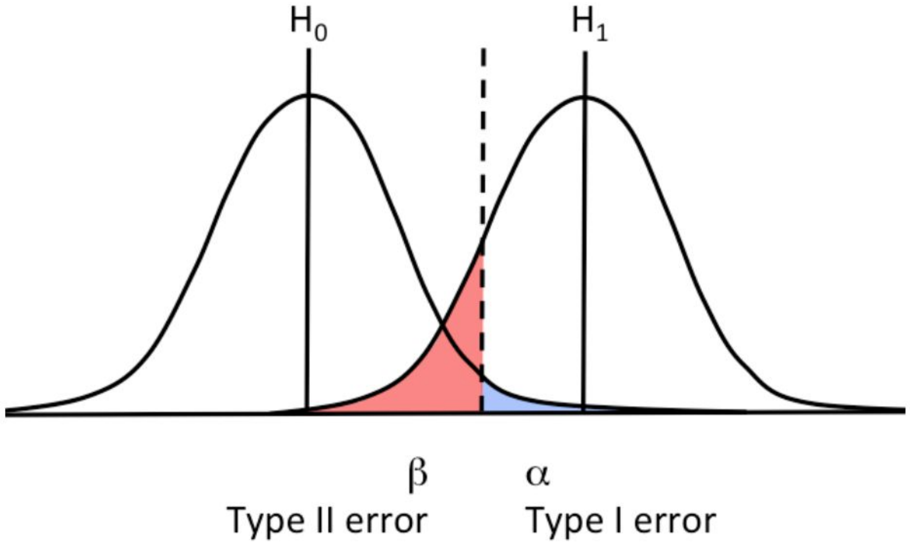
- **Type I:** Rejecting H_0 when it is true (e.g. saying men and women make different salaries on average when they don't)
- **Type II:** Failing to reject H_0 when it is false (e.g. saying their salaries are the same on average when they aren't)

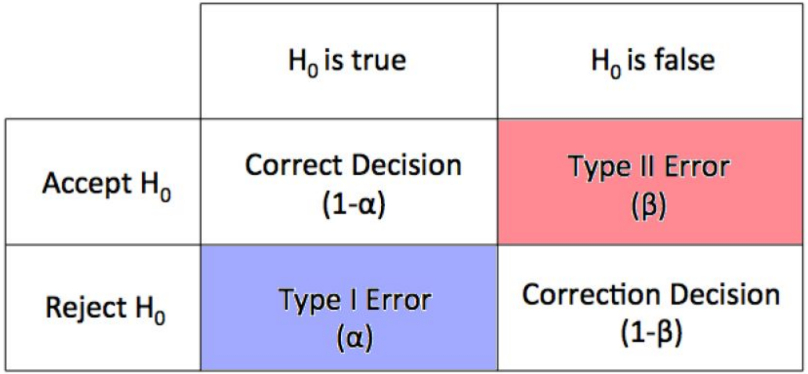
The **significance level (α)** is the amount of **Type I** error that we're willing to allow.

$\alpha = .05$ is a commonly chosen significance level, but it can vary (depending on the field)

Step 2: Choose a Significance Level

Sometimes it helps to think of whether we live in the H_O universe or the H_A universe.





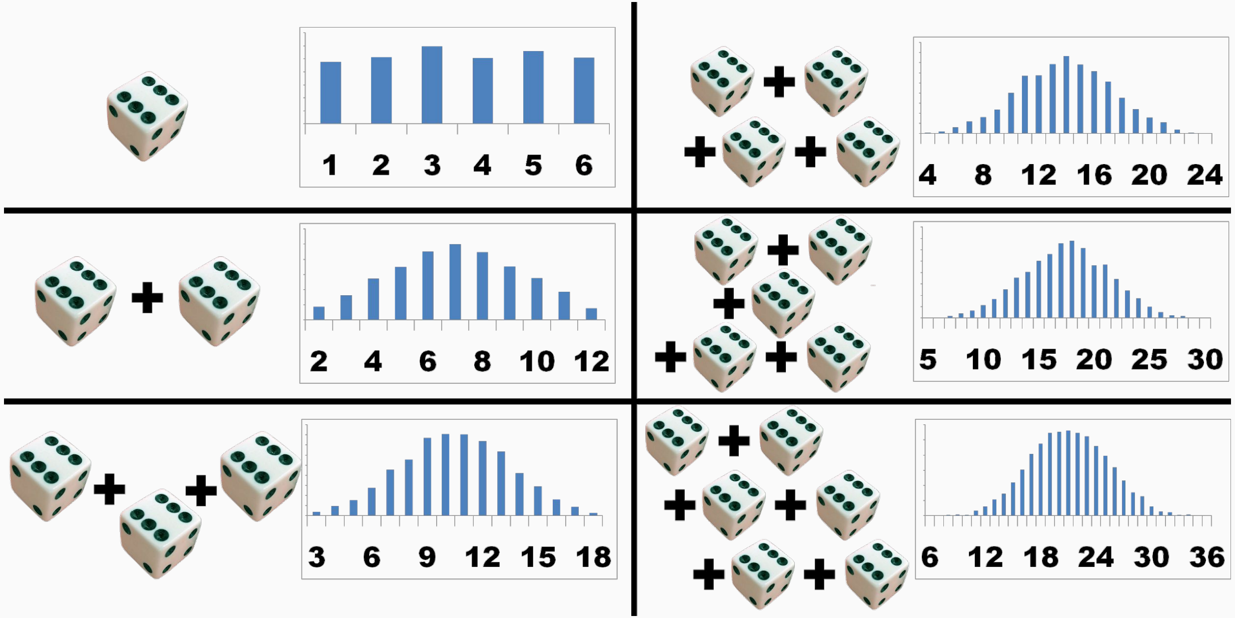
	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

False positive

False negative

Central Limit Theorem

Given certain conditions, the mean of a sufficiently large number of i.i.d. Random variables, will be approximately normal, regardless of the underlying distribution



→ This allows us to approximate the distribution for the H_0 universe and calculate a type I error for a given test statistic

Step 3: Compute Test Statistic

Often, the **test statistic** is based on a sample mean or comparison of sample means.

Examples:

- If we’re testing if a coin is fair, we could perform n trials, calculate the mean outcome (Heads=1, Tails=0), and compare to the expected mean (.5).
- If we’re wondering if caffeine makes students perform better on tests, we could compare mean test scores from a caffeinated group and a control group.

The test-statistic is generally stated in **units of standard deviation** or standard error relative to the distribution associated with the null hypothesis.

Step 3: Compute Test Statistic

The test-statistic is generally stated in **units of standard deviation** or standard error relative to the distribution associated with the null hypothesis.

$$\text{test statistic} = \frac{\textit{estimate} - \textit{value we hypothesize}}{\textit{standard error}}$$

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

From Central Limit Theorem

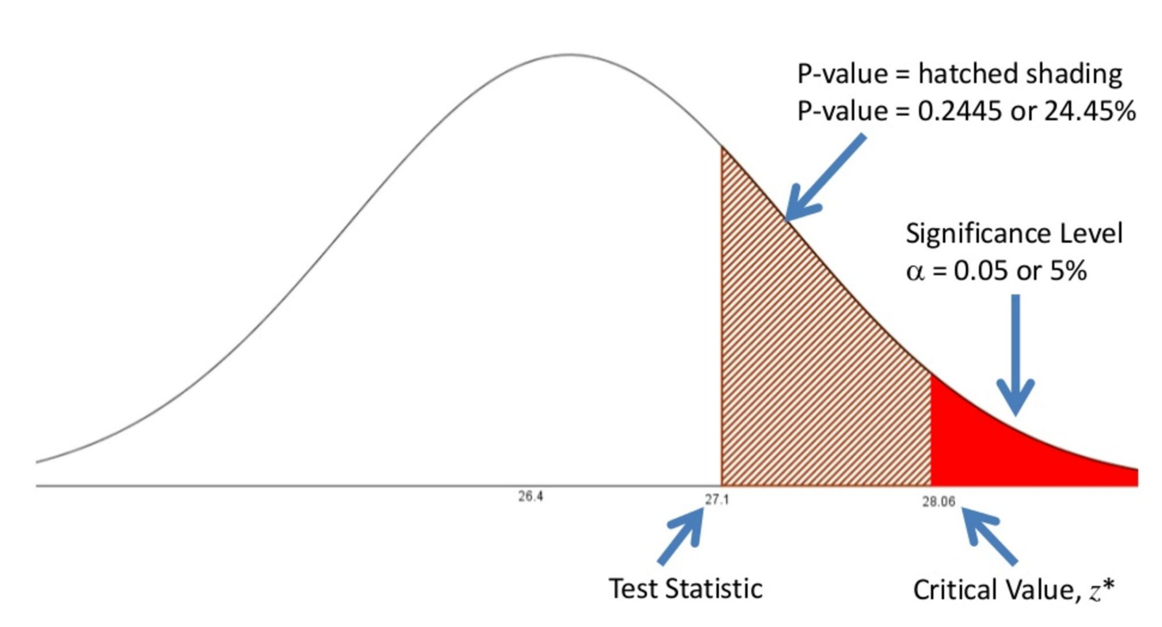
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Step 4: Calculate P-Value

The **p-value** is the probability of getting the observed result (or more extreme) given that the null hypothesis is true.

We usually use the t-distribution for this because we estimate the population variance by the sample variance, which for small samples is a bad estimate

The t-distribution has the parameter degrees of freedom which adjusts for that



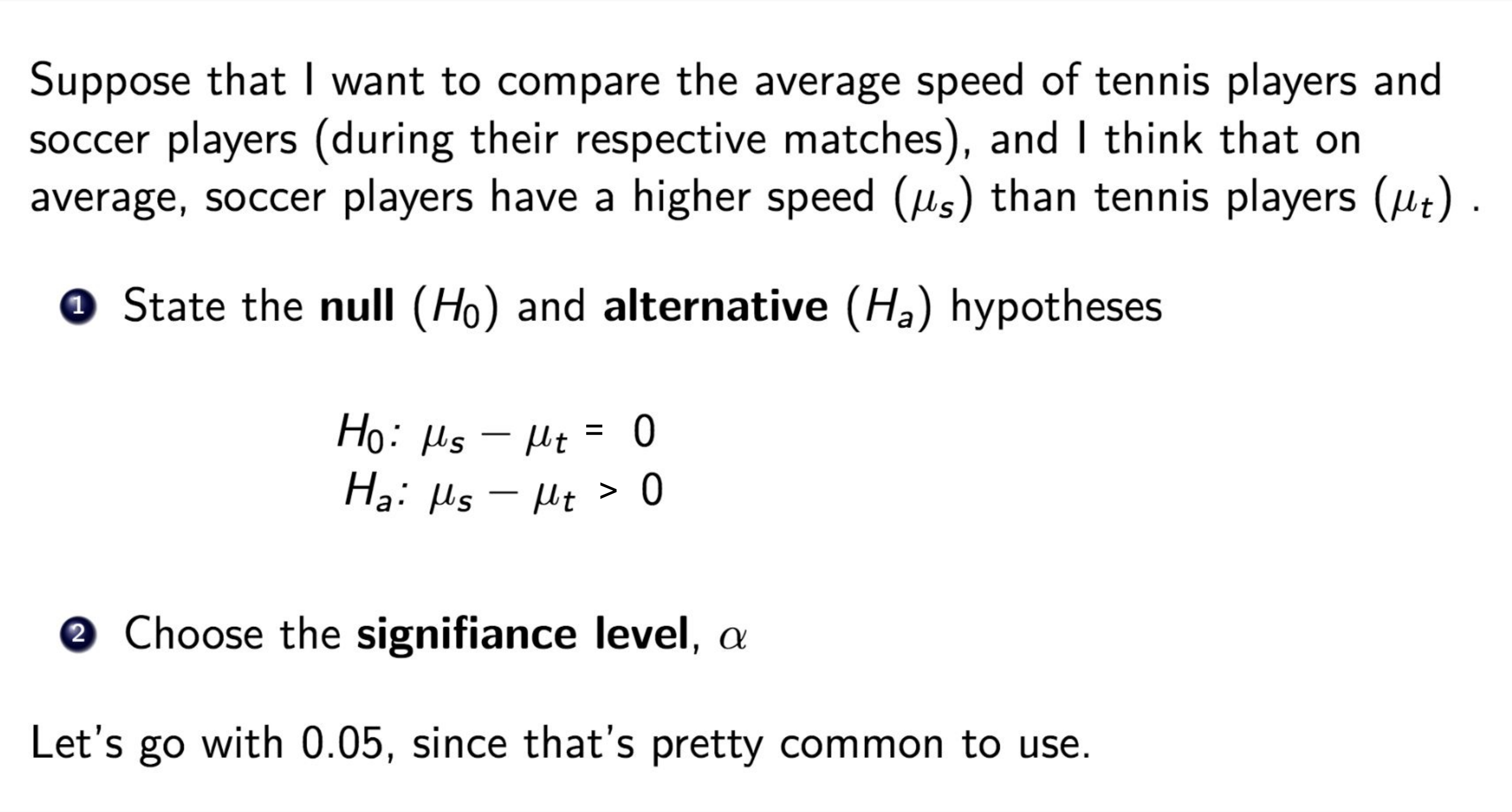
P-value = hatched shading
P-value = 0.2445 or 24.45%

Significance Level
α = 0.05 or 5%

Test Statistic

Critical Value, z^*

Example: Two Sample T-Test



Suppose that I want to compare the average speed of tennis players and soccer players (during their respective matches), and I think that on average, soccer players have a higher speed (μ_s) than tennis players (μ_t) .

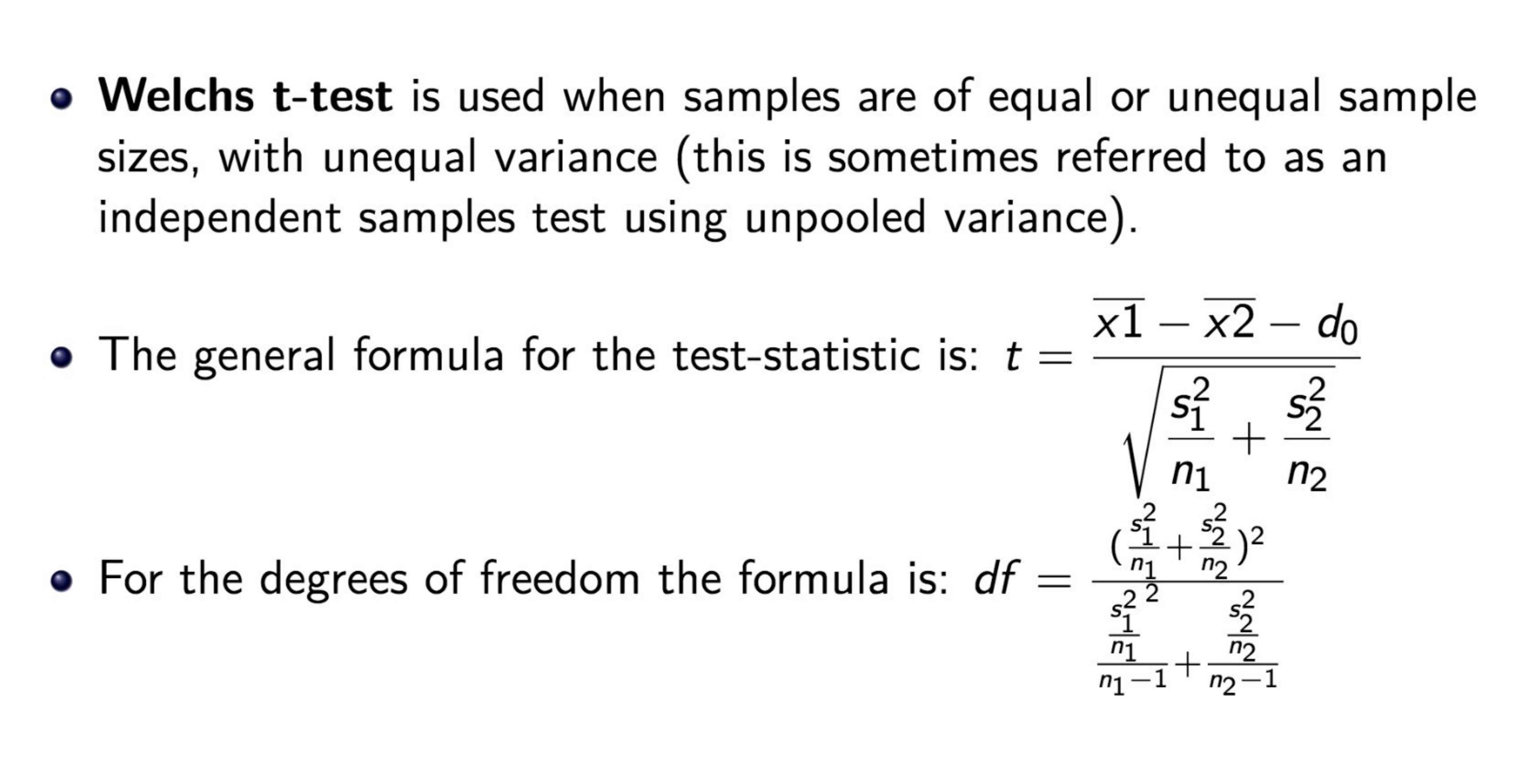
1. State the **null** (H_0) and **alternative** (H_a) hypotheses

$$H_0\colon \mu_s - \mu_t = 0$$
$$H_a\colon \mu_s - \mu_t > 0$$

2. Choose the **signifiance level**, α

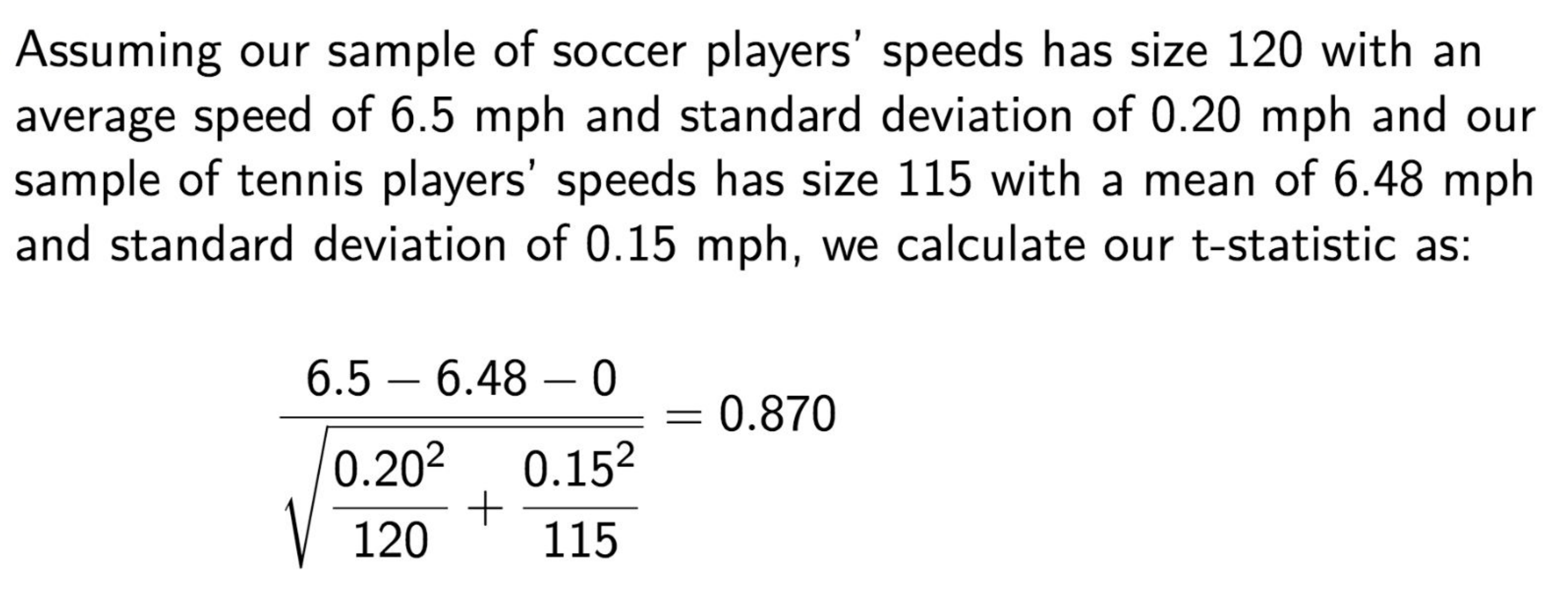
Let's go with 0.05, since that's pretty common to use.

Example: Two Sample T-Test



- **Welchs t-test** is used when samples are of equal or unequal sample sizes, with unequal variance (this is sometimes referred to as an independent samples test using unpooled variance).
- The general formula for the test-statistic is: $t = \frac{\overline{x1} - \overline{x2} - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- For the degrees of freedom the formula is: $df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{{\frac{s_1^2}{n_1}}^2}{n_1 - 1} + \frac{\frac{s_2^2}{n_2}}{n_2 - 1}}$

Example: Two Sample T-Test



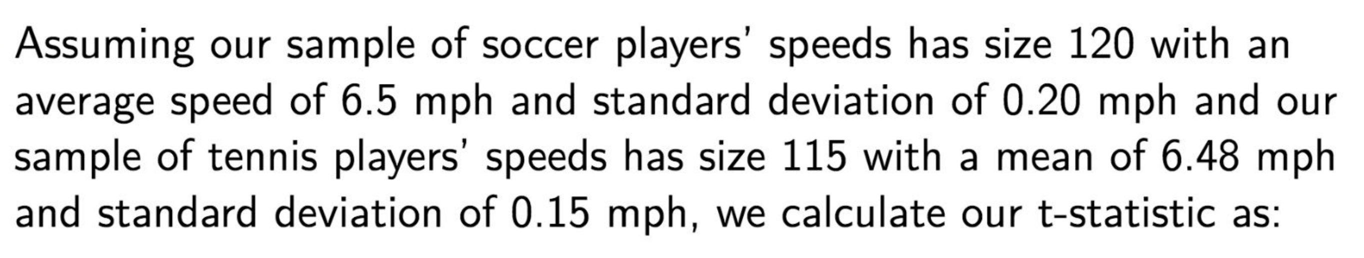
Assuming our sample of soccer players' speeds has size 120 with an average speed of 6.5 mph and standard deviation of 0.20 mph and our sample of tennis players' speeds has size 115 with a mean of 6.48 mph and standard deviation of 0.15 mph, we calculate our t-statistic as:

$$\frac{6.5 - 6.48 - 0}{\sqrt{\frac{0.20^2}{120} + \frac{0.15^2}{115}}} = 0.870$$

Using an old-fashioned t-table, we could look up a t-stat of .870 and find a p-value of 0.193.

Since this is above our .05 alpha we fail to reject the null and conclude both groups' speeds are not significantly different

Example: Two Sample T-Test in Python



Assuming our sample of soccer players' speeds has size 120 with an average speed of 6.5 mph and standard deviation of 0.20 mph and our sample of tennis players' speeds has size 115 with a mean of 6.48 mph and standard deviation of 0.15 mph, we calculate our t-statistic as:

But we have Python! Instead...

```
In [1]: mean_1, std_1, obs_1, mean_2, std_2, obs_2 = 6.5, .2, 120, 6.48, .15, 115

In [2]: scipy.stats.ttest_ind_from_stats(mean_1, std_1, obs_1, mean_2, std_2, obs_2,
equal_var = False)

Out[2]: Ttest_indResult(statistic=0.86957712916199181, pvalue=0.38547745845206016)

In [3]: _[1]/2

Out[3]: 0.19273872922603008
```

Because the default t-test is two-sided in scipy

Types of Test Statistics

There are many different types of test statistics and the appropriate one will vary. Some basic guidelines:

t-test	z-test
● Use when population variance is unknown ● **Small sample size (<30)** ● Built upon the t-distribution	● Use when population variance is known ● Unknown variance and **large sample size** ● Built on the normal distribution

For both t- and z- tests:

- **One-sample** tests are used when we want to compare a single population mean to a hypothesized value
- **Two-sample** tests are used to compare two population means with each other
- **One- vs two-sided:** Default in Python is two-sided, so divide p-value by 2 for a one-sided test

Test Statistics in Python

Tests below all return `[test-statistic, p-value]`:

Test	Description
`scipy.stats.ttest_1samp (values, mean)`	Calculates the T-test for the mean of ONE group of scores. (Test if the mean of the sample is different from specified mean)
`scipy.stats.ttest_ind (values_a, values_b, equal_var)`	Calculates the T-test for the means of *two independent* samples of scores.
`scipy.stats.ttest_ind_from_stats (mean1, std1, nobs1, mean2, std2, nobs2, equal_var)`	T-test for means of two independent samples from descriptive statistics.
`scipy.stats.ttest_rel (values_a, values_b)`	Calculates the T-test on TWO RELATED samples of scores, a and b. (Paired t-test, ex.: group of students retakes a test)
`statsmodels.stats.weightstats.ztest (x1, x2=None, value=0, alternative='two-sided', usevar='pooled', ddof=1.0)`	Calculates the T-test for mean based on normal distribution (Z test), one or two samples. In the case of two samples, the samples are assumed to be independent.

Example

⚠ STOP AND PRACTICE

Suppose you're trying to tell if giving students caffeine influences their test scores. Group A has 40 students who drink coffee and they score 93 with standard deviation of 10. Group B has 50 students who score 89 on average with an 8-point standard deviation. The following code generates test scores for each group: (Set random seed only so everyone in class has the same random arrays; must re-run before each statement if in ipython session.)

```
import numpy as np
np.random.seed(42)
coffee_group = np.random.normal(93, 10, 40)
np.random.seed(42)
control_group = np.random.normal(89, 8, 50)
```

Determine if your caffeine influences test scores.

```
In [1]: scipy.stats.ttest_ind(coffee_group, control_group, equal_var=False)
Out[1]: Ttest_indResult(statistic=1.9660382290361424, pvalue=0.053109281349055369)
```

Example

⚠ STOP AND PRACTICE

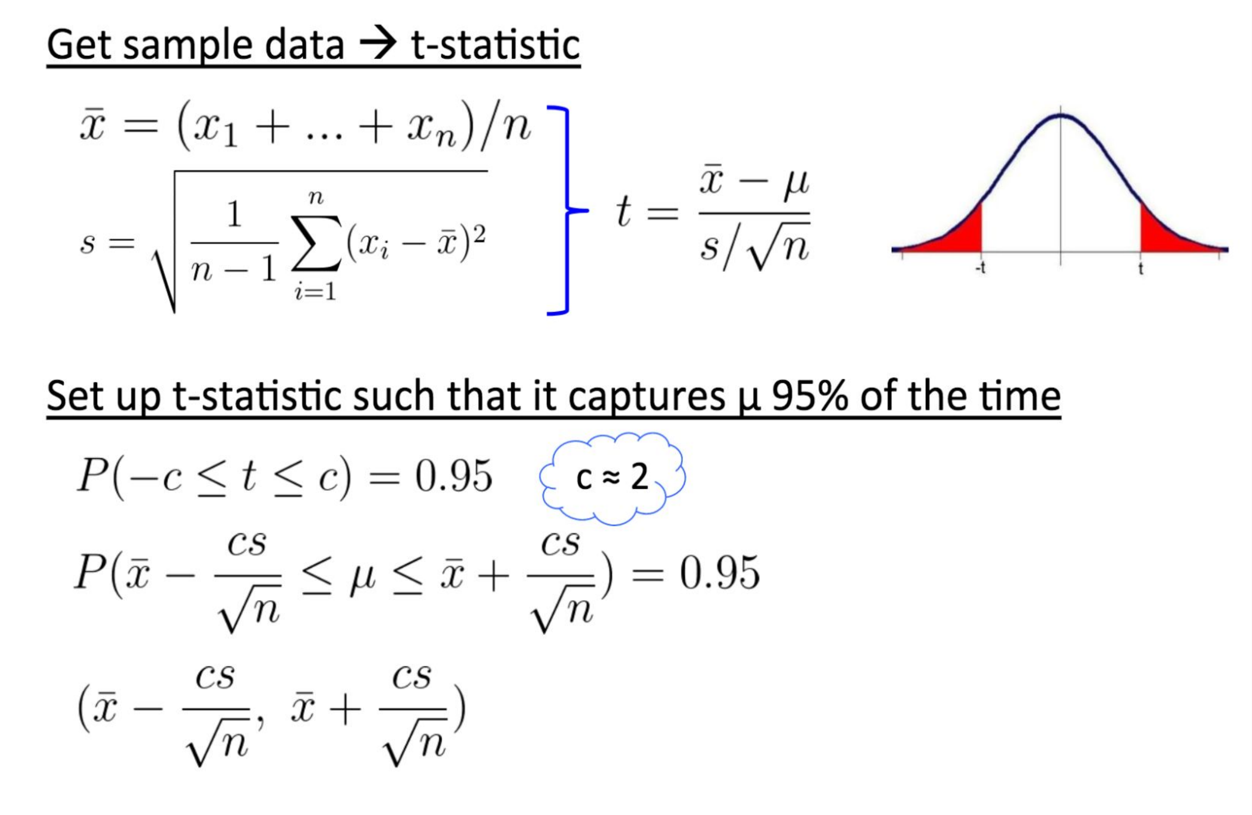
Suppose you own 120 kegs of beer and the industry-standard is 15.5 gallons per keg. Assume you measure all 120 and find that the mean volume of your kegs is 15.4 gallons with a standard deviation of 0.2 gallons, and `keg_list` is the list of each keg's volume:

```
import numpy as np
np.random.seed(42)
keg_list = np.random.normal(15.4, .2, 120)
```

Determine if your inventory of kegs have less volume than what the beer industry wants you to think.

```
In [1]: scipy.stats.ttest_1samp(keg_list, 15.5)[1]/2
Out[1]: 1.6758414241350826e-10
```

Constructing a confidence interval



Get sample data → t-statistic

$$\left.\begin{aligned} \bar{x} &= (x_1 + \ldots + x_n)/n \\ s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}\right\} t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Set up t-statistic such that it captures μ 95% of the time

$$P(-c \leq t \leq c) = 0.95$$

c ≈ 2

$$P(\bar{x} - \frac{cs}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{cs}{\sqrt{n}}) = 0.95$$

$$(\bar{x} - \frac{cs}{\sqrt{n}},\ \bar{x} + \frac{cs}{\sqrt{n}})$$

Say “With a confidence level of 95% the true population mean lies between a and b”

Step 2: Choose a Significance Level

⚠ **STOP AND THINK**

Imagine you conduct an hypothesis test with a $\alpha = .05$ and repeat it 20 times.

How might this be problematic?

Multiple comparisons

We run the first test with significance level 0.05

→ Probability of not getting a Type I error: 0.95

We run the test a second time with significance level 0.05

→ Probability of not getting a Type I error: $0.95^2 = 0.9025$

… after 20 tests: Probability of not getting a Type I error: $0.95^{20} = 0.36$!!

Example: We want to test 100 variants of our website and run 100 tests. With a significance level of 0.05 we would expect 5 of them to produce a "success", even if there no significant changes!

Multiple comparisons

There are many ways to correct for this (even though many people forget to correct for it at all!)

A popular one is the “Bonferroni” correction, which simply adjusts the desired significance level for each test:

$$\alpha_B = \alpha/m$$

5 Steps of Hypothesis Testing

1. State the **null hypothesis** (H0) and **alternative hypothesis** (HA)
2. Choose a **significance level** (α), typically .05
3. Select statistical test, and compute appropriate **test statistic**
4. Compute **p-value** based on test statistic
5. State **conclusion**
    - If p-value < α: Reject H0
    - If p-value > α: Fail to reject H0

χ^2-test and experimental design

Objectives

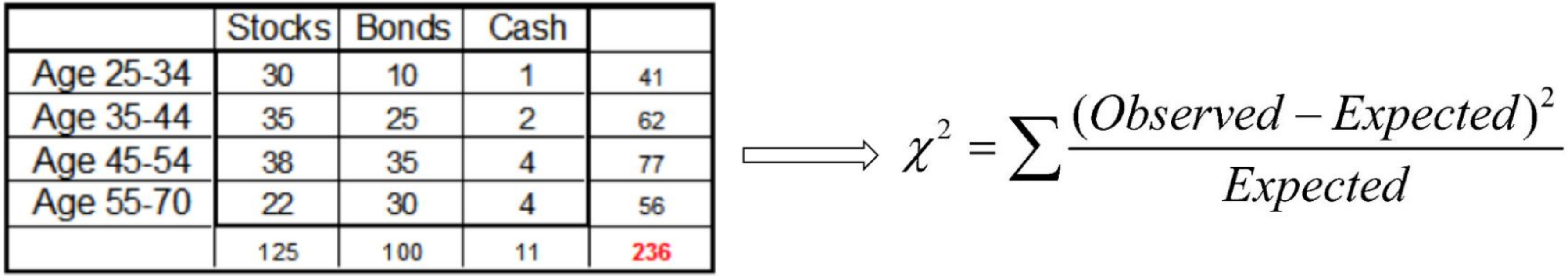
- Know when to perform a χ^2-test
  - Goodness of fit
  - Independence
- Perform a χ^2-test
- Observational vs experimental studies and experimental design problems
- Be aware of confounding factors

galvanize

χ^2-test

Similarly to the hypothesis testing seen this morning we are trying to compare measured values to some hypothetical values and decide if they are different

The χ^2-test is generally used for comparing a set of categorical data to a hypothesized distribution for that data



	Stocks	Bonds	Cash	
Age 25-34	30	10	1	41
Age 35-44	35	25	2	62
Age 45-54	38	35	4	77
Age 55-70	22	30	4	56
	125	100	11	236

$$\Longrightarrow \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Is there a relationship between age and investment preference?

χ^2-distribution

If Z_1, … Z_k are independent, unit variance normally distributed random variables then the sum of their squares follows a χ^2-distribution with k degrees of freedom

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi^2_k.$$

with

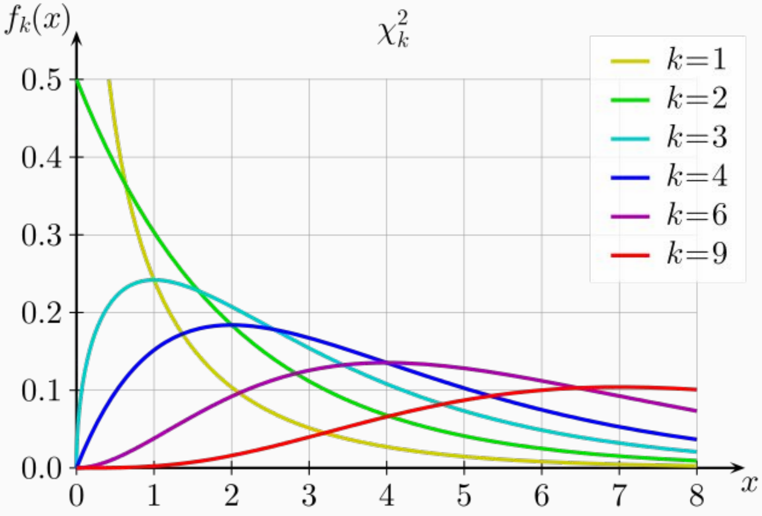
$$Q = \sum_{i=1}^{k} Z_i^2,$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

follows a chi-squared distribution, because

$$\frac{(O_i - E_i)}{\sqrt{E_i}}$$

is approximately normal with unit variance



Example: fair dice (Goodness of fit)

I throw a die 60 times and count the number of outcomes

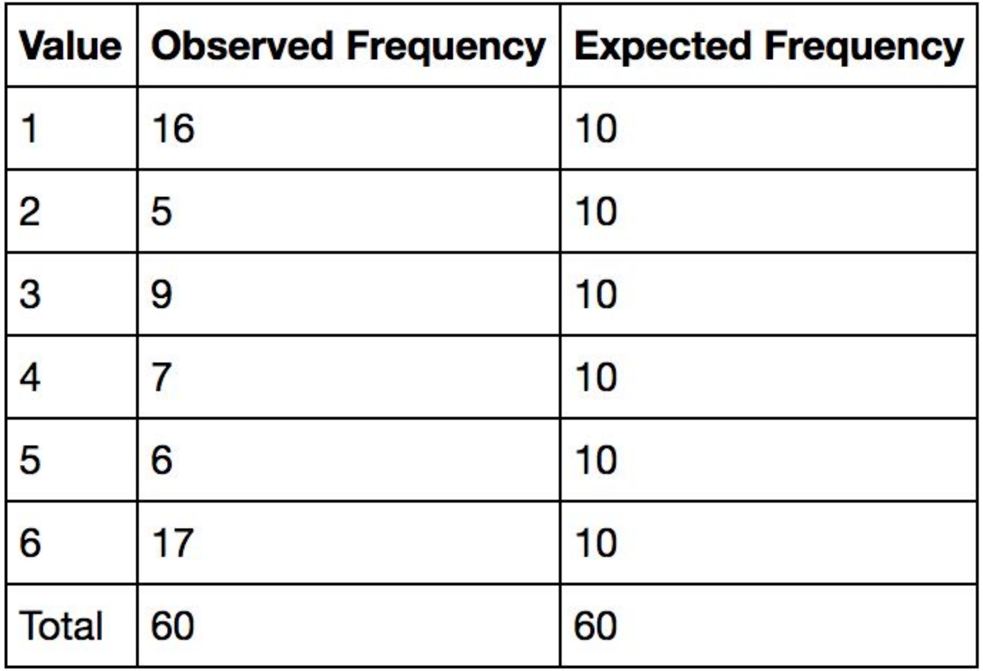
What is the expected distribution?

Uniform!

```
obs_table = np.array([16,5,9,7,6,17])
exp_table = np.array([10,10,10,10,10,10])
chi2_stat = sum((exp_table -
obs_table)**2/exp_table)
print "Chi2 Statistic: {}".format(chi2_stat)
print "Critical Chi2 Value:
{:0.2f}".format(stats.chi2.ppf(0.95,5))
```

```
Chi2 Statistic: 13.6
Critical Chi2 Value: 11.07
```

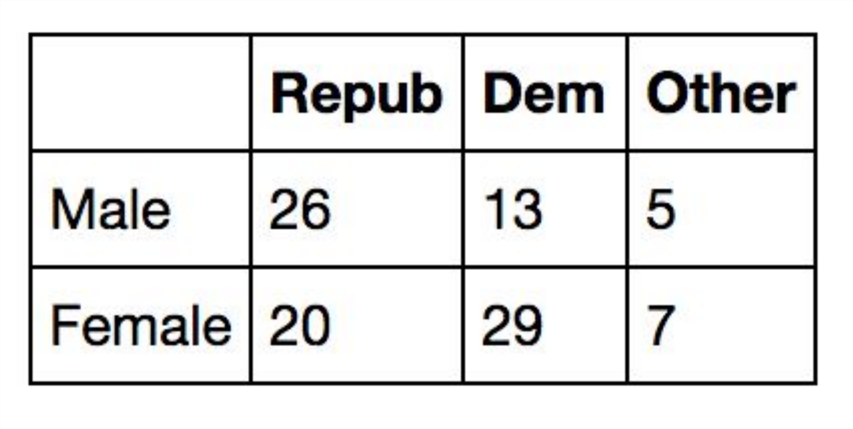
Or: `stats.chisquare(obs_table,exp_table)`



Value	Observed Frequency	Expected Frequency
1	16	10
2	5	10
3	9	10
4	7	10
5	6	10
6	17	10
Total	60	60

Example: Voting and gender (test for independence)

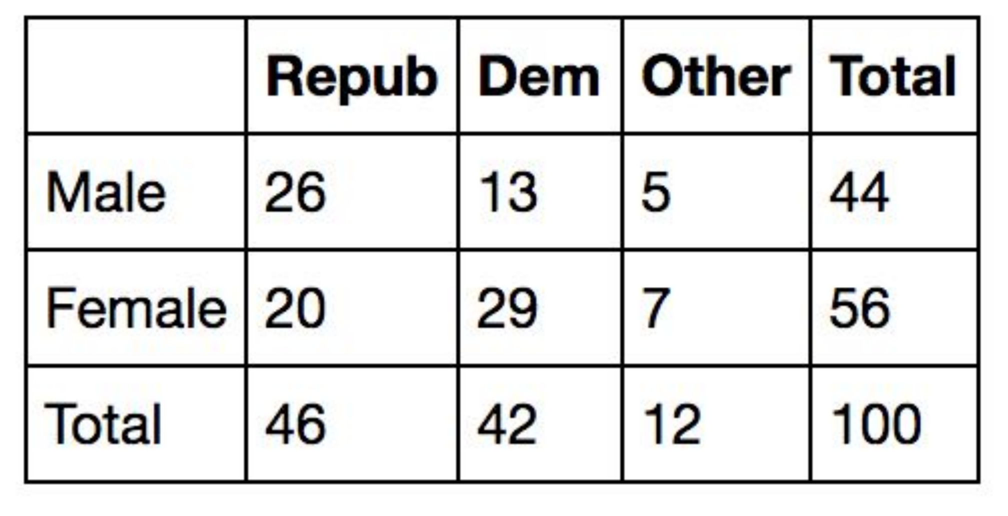
Is voting behavior independent of gender? (or, similarly, is churn behavior independent of operating system?)



	Repub	Dem	Other
Male	26	13	5
Female	20	29	7

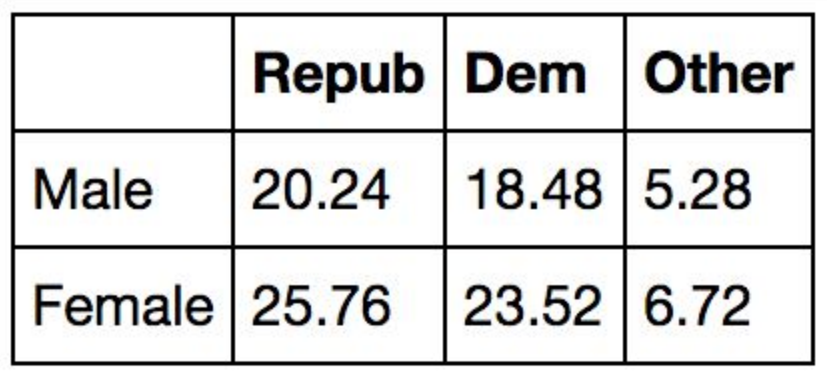
Example: Voting and gender (test for independence)

If independent of gender, we assume the distribution within each gender to be insignificantly different from the distribution of the pooled genders



	Repub	Dem	Other	Total
Male	26	13	5	44
Female	20	29	7	56
Total	46	42	12	100

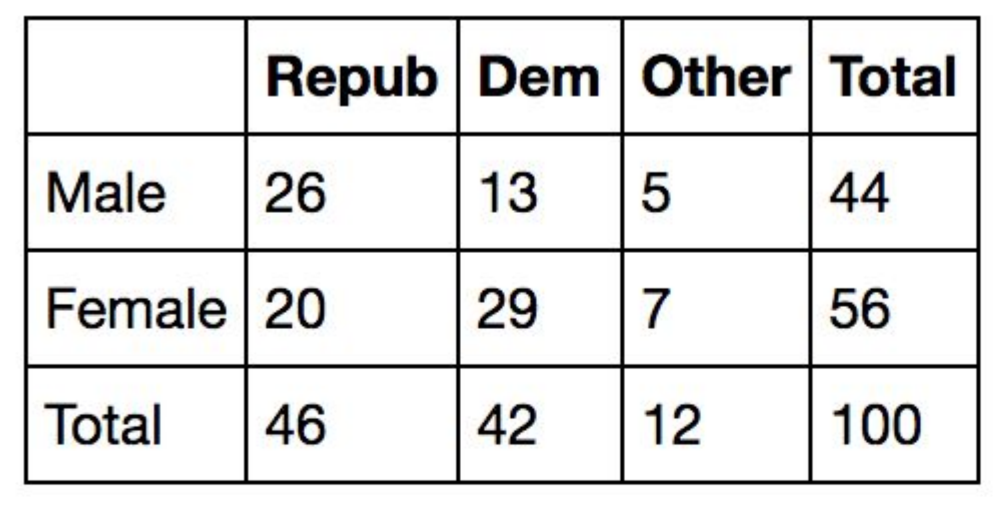
→ calculate expected values from pooled values



	Repub	Dem	Other
Male	20.24	18.48	5.28
Female	25.76	23.52	6.72

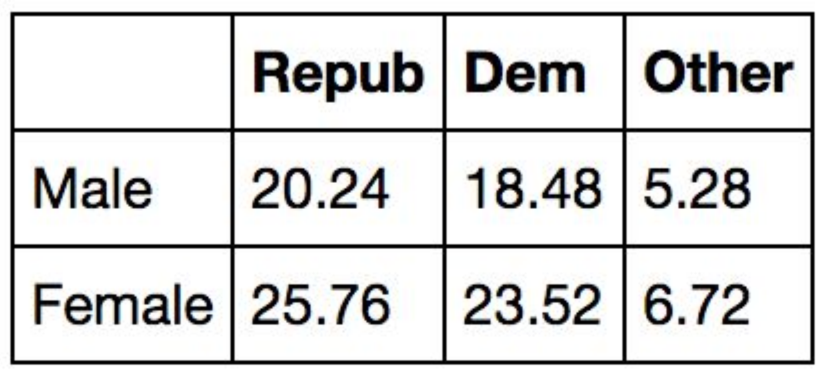
Example: Voting and gender (test for independence)

Observed:



	Repub	Dem	Other	Total
Male	26	13	5	44
Female	20	29	7	56
Total	46	42	12	100

Expected:



	Repub	Dem	Other
Male	20.24	18.48	5.28
Female	25.76	23.52	6.72

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Example: Voting and gender (test for independence)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

χ^2 = (26-20.24)^2/20.24 + (13 - 18.48)^2/18.48 + … = 5.86
(critical value for significance level 0.05 would be 5.991 according to Chi-square table with dof = 2 #[(columns-1)*(rows-1)]

Scipy:

```
cont_table = np.array([[26,13,5],[20,29,7]])
chi2,pval,dof,exp_array = stats.chi2_contingency(cont_table)
```

```
Chi2 Value: 5.85549931435
p-value: 0.0535173350908
dof: 2
Expected Value Array:
[[ 20.24  18.48   5.28]
 [ 25.76  23.52   6.72]]
```

Experimental vs observational studies

Experimental

- Randomly assign groups to minimize confounding
- **Apply treatments** to experimental units and observe the effect of treatments
- Can be used to **establish causality**
- Ex: Randomly assign homework to students and measure performance in class

Observational

- Observe subjects and measure variables of interest **without assigning treatments** to subjects (self-selection)
- Confounding factors very likely and require adjustment
- **Cannot be used to establish causality**
- Ex: observe grades of students who did and didn’t do their homework

Confounding factors

A confounding factor is an extraneous factor that correlates with the dependent variable (performance) and the independent variable (homework)

What could the confounding factor be?

- How hard-working a student is
- Hard-working students tend to do their homework AND have better grades, they aren’t better BECAUSE they do their homework!

Experimental design

Removing all confounding factors is close to impossible

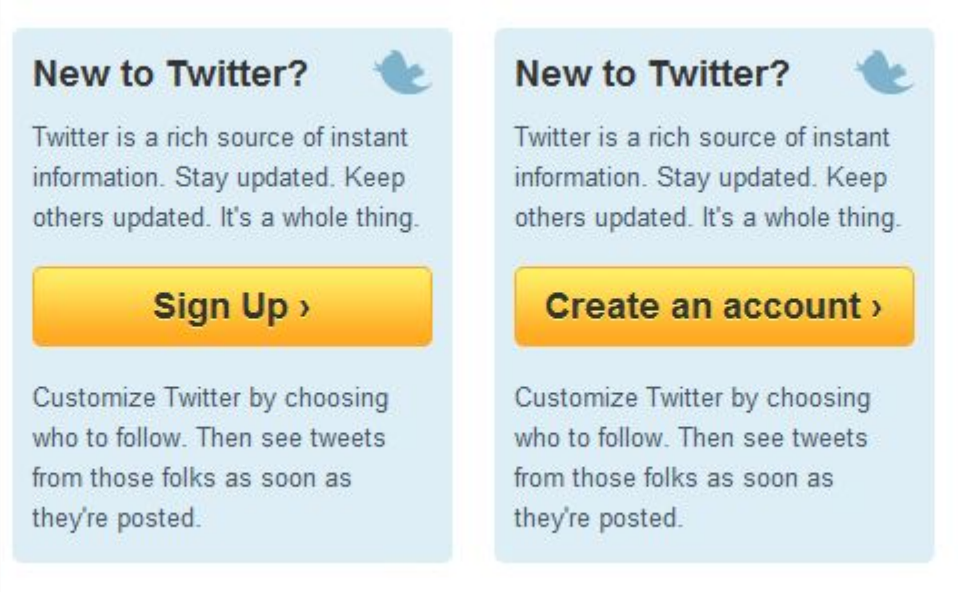
Do the best you can!

- Randomization into groups of equal size:
  - Randomly generate number from 0 to 1 for each student
  - Assign homework if > 0.5
- Try to maintain independence
  - Students shouldn’t know if other students have homework or not
  - If they know, it might affect their performance
- If possible (e.g. medical treatment) maintain blindness
  - Patients should not know if they get placebo or not
  - Double-blind: experimenter doesn’t know who is assigned to what group

Questions

- What are the general steps of a hypothesis test?
- How do we test the difference of two means?
- Why does this work even if compared populations are very non-normal?
- When do we use a chi-squared test?
- What is the multiple comparison problem and how do we correct for it?
- What is a p-value? Type I error? Type II error?

A/B testing



New to Twitter?

Twitter is a rich source of instant information. Stay updated. Keep others updated. It's a whole thing.

Customize Twitter by choosing who to follow. Then see tweets from those folks as soon as they're posted.

New to Twitter?

Twitter is a rich source of instant information. Stay updated. Keep others updated. It's a whole thing.

Customize Twitter by choosing who to follow. Then see tweets from those folks as soon as they're posted.