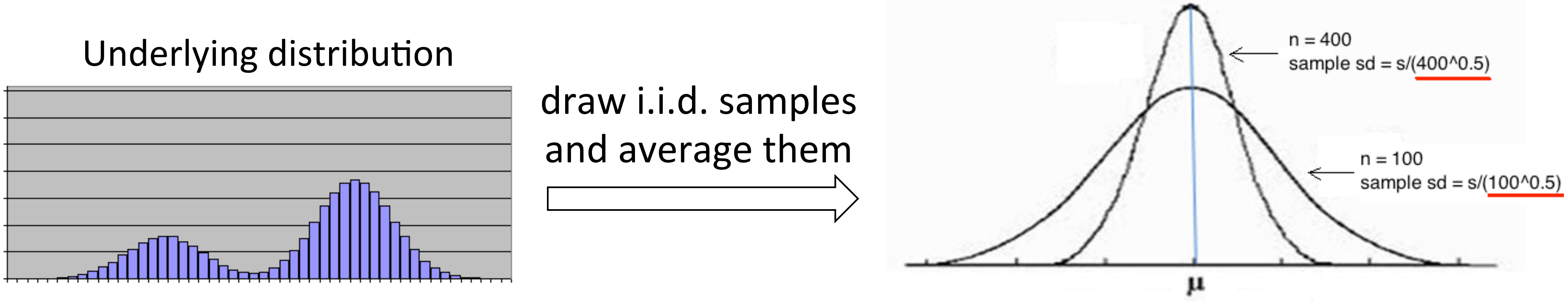
Hypothesis Testing

Overview

- Hypothesis Testing Framework
- Two-sample t-test for Comparison of Means
- Two-sample z-test for Comparison of Proportions
- Multiple Comparisons Problem
- Chi-Square Test
  - Links to test of Comparison of Proportions
- Experimental Design

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.


Underlying distribution

draw i.i.d. samples and average them

Central Limit Theorem

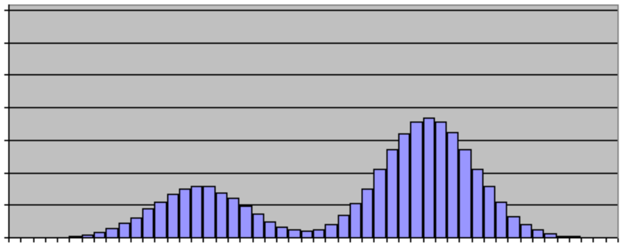
- Not only is the sample mean normally distributed, we have….

$$\boxed{\bar{X} \sim Normal(u, \frac{\sigma^2}{n})}$$

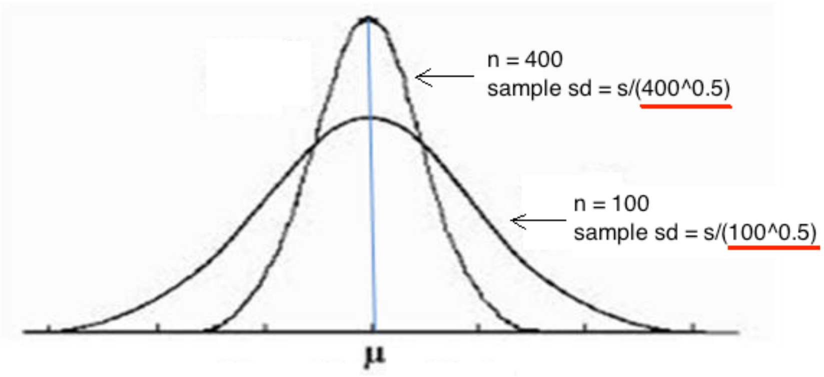
- And as usual, from any normally distributed random variable, we can derive a standard normal variable. In this case…

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Underlying distribution



draw i.i.d. samples and average them



Hypothesis Testing – Framework

- Interested in some population parameter(s), but don't have access.
  - But…you can get some sample data!
- And, you have some competing hypotheses to test using sample data
  - H_0 (Null): Men and Women make the same amount of money
  - H_A (Alternative): Men make more money than Women do
- The Setup: **Innocent (H_0) until proven Guilty (H_A)**
  - In other words, we suppose the null hypothesis to be true, and look for evidence to the contrary.

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."

Hypothesis Testing – Framework

1. State the null hypothesis (H_0)
   and the alternative hypothesis (H_A)
2. Choose significance level, α, typically α=0.05.
3. Select statistical test, and compute appropriate test statistic
4. Compute p-value based on test statistic from step 3.
   - If p-value < α => Reject H_0 in favor of H_A
   - If p-value > α => Fail to reject H_0

Hypothesis Testing – Framework

1. State null hypothesis (H_0) and alternative hypothesis (H_A)

   H_0: $\mu = 100$

   H_A: $\mu \neq 100$

2. Choose significance level, α

   $\alpha = 0.05$

3. Compute appropriate test statistic using collected data.

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

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

⟹

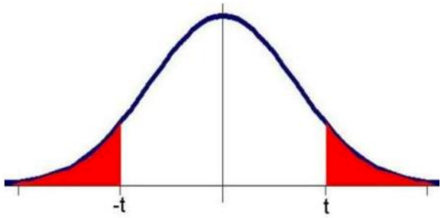
n=30

Sample mean ($\bar{\mathbf{x}}$)= 102

Sample standard deviation (*s*) = 7

t = (102-100)/(7/(30^.5)) = **1.565**

4. Compute p-value based on test statistic



Since we are doing a two-sided test, we take area to left of t = -1.565 and right of t = +1.565

p-value ≈ 0.1284 > 0.05 ⟹ Fail to reject null H_0

Hypothesis Testing – Framework

- Notice we "fail to reject the null", H_0: $\mu = 100$
  - We don't conclude that $\mu = 100$!
- What if all sample stats were the same, except instead, n = 100??

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

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

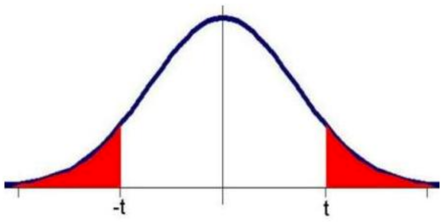
⟹

n=**100**

Sample mean ($\bar{\mathbf{x}}$)= 102

Sample standard deviation (*s*) = 7

t = (102-100)/(7/(**100**^.5)) = **2.857**

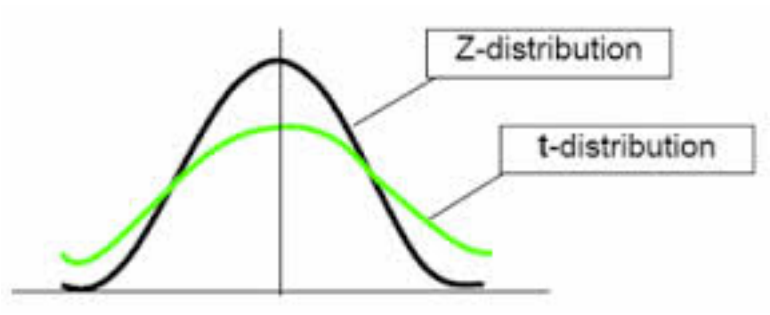


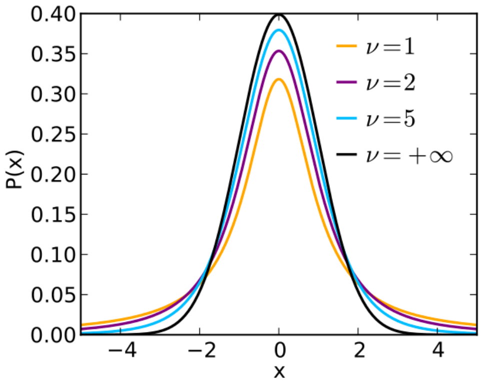
Since we are doing a two-sided test, we take area to left of t = -2.857 and right of t = +2.857

p-value ≈ 0.0052 < 0.05 ⟹ Reject null H_0 in favor of H_A

Hypothesis Testing – Framework

- What is a t-distribution? fat-tailed Normal, that approaches normal as d.f. → ∞

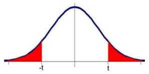




- Why on earth would $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ follow a t-distribution?

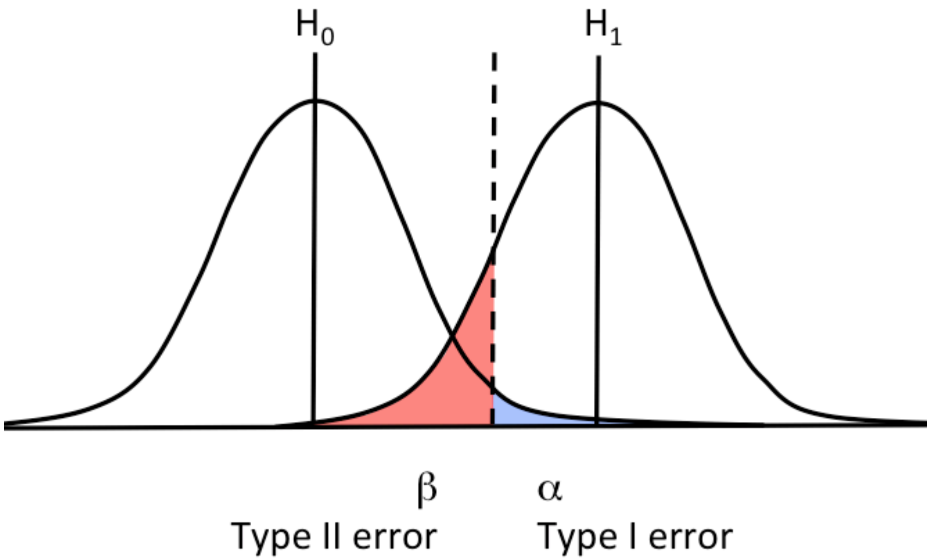
  Central Limit Theorem!

Type I and Type II Errors

- p-value is **“the probability of observing the data we observed, or more extreme, given the null hypothesis is true”**
- Conceptually, we simply look at the tail end(s) of what we might expect for the distribution of the sample mean, under the null hypothesis. But we could always in earnest, be wrong, due to random variation. We’re willing to accept this α of the time. 

	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)

Type I and Type II Errors



	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)

What happens when we increase the sample size?

A note on confidence intervals

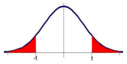
Setup

H_0: $\mu = 100$
H_A: $\mu \neq 100$
Chose significance level, $\alpha = 0.05$

Sample Data

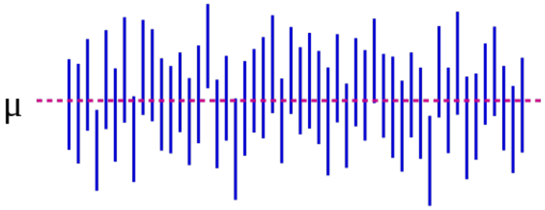
n=30
Sample mean ($\bar{\mathbf{x}}$)= 102
Sample standard deviation (*s*) = 7

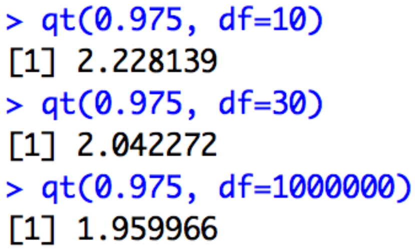
t = (102-100)/(7/(30^.5)) = 1.565
p-value ≈ 0.1284 > 0.05



Construct Confidence Interval for population mean, μ

$$\boxed{(\bar{x} - t_{\alpha/2} * \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2} * \frac{s}{\sqrt{n}})}$$





```
> qt(0.975, df=10)
[1] 2.228139
> qt(0.975, df=30)
[1] 2.042272
> qt(0.975, df=1000000)
[1] 1.959966
```

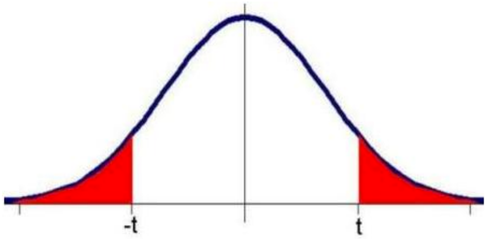
102 +/- **2***[**7/(30^.5)**] = (99.44, 104.55)

- Don't say "The probability that true population mean, μ, is in range (88.28, 115.72) is 95%"
- Say "With confidence level 95%, μ lies in the interval (88.28, 115.72)"

A note on confidence intervals

<u>Get sample data → t-statistic</u>

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



<u>Set up t-statistic such that it captures μ 95% of the time</u>

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

c ≈ 2

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

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

Quick Recap

Hypothesis Testing Framework

1. Setup: H_0 and H_A, α
2. Compute test statistic and compare to appropriate distribution
   - Often, test statistic is a sample mean (or difference of two sample means), so we compare it to the t-distribution, a gift from the **CLT**!
3. Compute p-value based on previous step
   - If p-value < α => Reject H_0 in favor of H_A
   - If p-value > α => Fail to reject H_0

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

Two-sample comparison of means, One-sample proportion test,
Two-sample comparison of proportion…all pretty much the same idea!

Two-sample Comparison of Means

1. State null hypothesis (H_0) and alternative hypothesis (H_A), let α=0.05

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 > \mu_2$

2. Compute appropriate test statistic using collected data.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

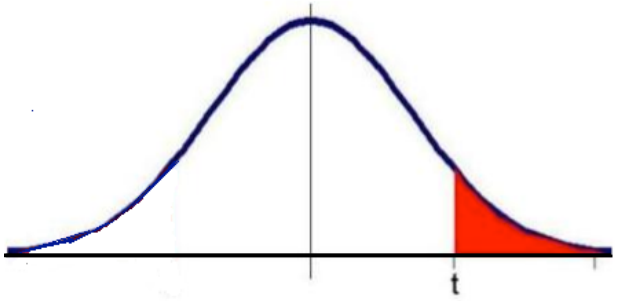
where $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

⟹

n_1=20, n_2=30
$\bar{\mathbf{x}}_1$= 101, $\bar{\mathbf{x}}_2$= 95
s_1=7, s_2=5

t = (101-95)/((49/20+25/30)^.5) = **3.311**

4. Compute p-value based on test statistic



Since we are doing a **one**-sided test, we take area to right of t = +3.311

p-value ≈ 0.00116 < 0.05 ⟹ Reject null H_0 in favor of H_A

Two-sample Comparison of Means

Assumptions

- Population distributions normal (often not true)
  - CLT to the rescue! Helps if…
    - Population Distributions are close to normal
    - Sample n-size is large

- Standard Deviations are equal (often not true)
  - We applied "Welch's t-test", which works with equal or unequal sample sizes, and unequal variances
  - There are many variations
    - Equal sample sizes, equal variance
    - Equal or unequal sample sizes, equal variance
    - Equal or unequal sample sizes, unequal variance

Two-sample Comparison of Proportions

1. State null hypothesis (H_0) and alternative hypothesis (H_A), let α=0.05

   H_0: $p_1 = p_2$

   H_A: $p_1 > p_2$

2. Compute appropriate test statistic using collected data.

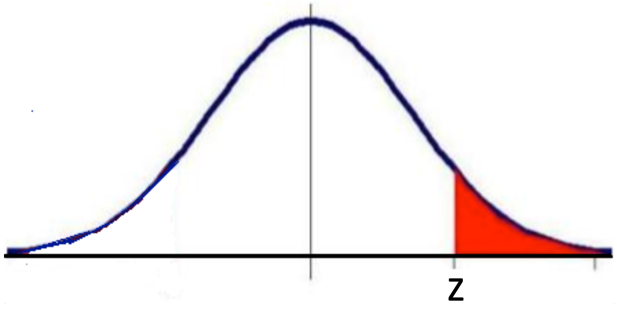
$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

⟹

n_1=300, n_2=1000
$\hat{p}_1$ = 0.05, $\hat{p}_2$ = 0.03

z = (0.05-0.03)/
((0.05*0.95)/300+(0.03*0.97)/1000)^0.5
= 1.46

3. Compute p-value based on test statistic



Since we are doing a **one**-sided test, we take area to right of z = +1.46

p-value ≈ 0.072 > 0.05 ⟹ Fail to reject null H_0

Two-sample Comparison of Proportions

How's this different from hypothesis testing for means? Not much.

- Really, we're still averaging X_1, X_2, …, X_n, except instead of taking on any values, they take on only 0 or 1.

Generally: $X \sim ?(\mu, \sigma^2) \xrightarrow{\text{by CLT}} \bar{X} \sim \ Normal(\mu, \frac{\sigma^2}{n})$

Here: $X \sim Bernoulli(p) \xrightarrow{\text{by CLT}} \bar{X} \sim Normal(p, \frac{p(1-p)}{n})$

Why z?

- t-test was due to estimating σ with sample sd, s.
- Here instead of (μ,σ) we just have a single parameter, p.

Multiple Comparisons Problem

Say we want to test a hypothesis, and set α=0.05

- 1st time we run a test, 5% chance of getting Type I Error
  ⬄ 95% chance of not getting a false positive.
- 2nd time we run a test, also 5% chance of Type I Error.
  - Probability that no Type I Error for both tests is 0.95^2 = 0.9025
- …after n tests
  - Probability of no Type I Error for any of n tests is **0.95^n**

Ex. Suppose we try 100 variations of original layout of a website with small tweaks such as magenta button color, panda icon, etc.

- Even if all changes made no difference, expect ~5 variations to be “successful”.
- There are many methods to account for this
  - **Bonferroni** adjustment, Fisher’s least-significant-difference, Duncan’s test, Scheffé’s test, Tukey’s test, and Dunnett’s test

Multiple Comparisons Problem – Bonferroni Correction

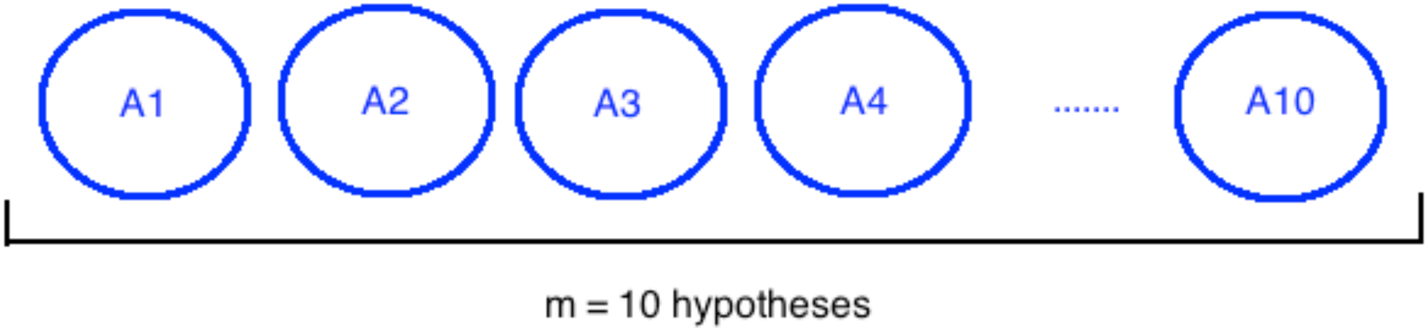
- Hypotheses $H_1, \ldots, H_m$ with corresponding p-values $p_1, \ldots, p_m$
- We know that $\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$ Where A_i is any event
- So we have family-wise error rate bounded by…

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha$$

In short, we're safe if we use **α/m** instead of α, when we examine our resulting p-values, $p_1, \ldots, p_m$

Multiple Comparisons Problem – Bonferroni Correction

- In pictures…suppose we test **10** hypotheses, so that m=10

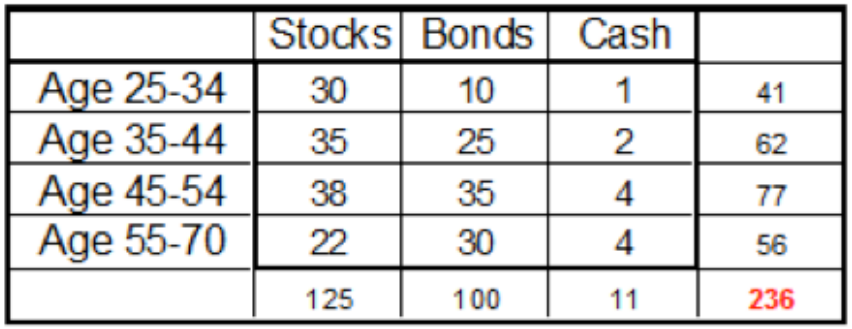


- If α = 0.05, we want overall Type I error to be bounded by 5%
- In the worst case scenario, our tests are independent (having nothing to do with each other). It's then conservative to measure each hypothesis against an adjusted significance level, **0.05/10 = 0.005.**

Afternoon

Chi-Square Test

- General method for comparing fact with theory
  - Pretty broad huh?
- Approach assumes sampled units fall randomly into cells, and that chance of a unit falling into particular cell can be estimated from the theory we're testing (not unlike H_0)
  - Similar. Assume something (H_0), collect some data, see if test statistic leads one to want to reject that assumption.



	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

Is there a relationship between age and investment preference?

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

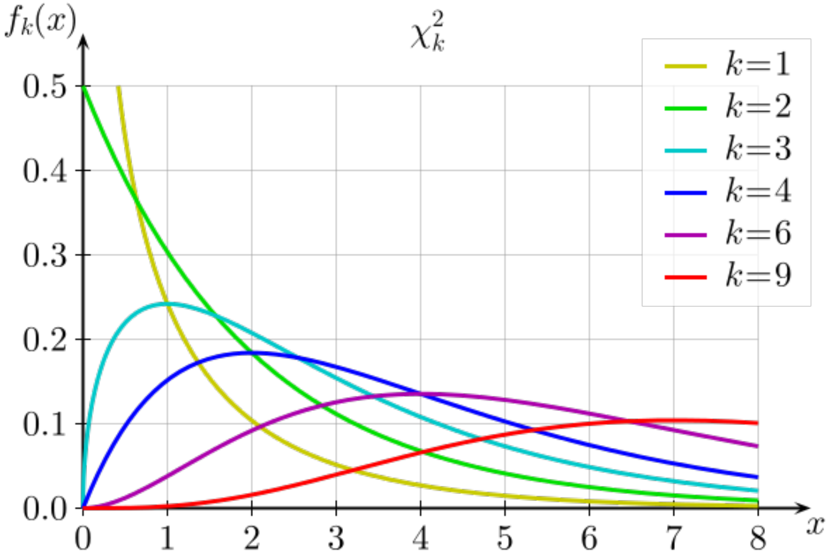
Chi-Square Distribution

If Z_1, ..., Z_k are independent, standard normal random variables, then the sum of their squares,

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

is distributed according to the **chi-squared distribution** with *k* degrees of freedom. This is usually denoted as

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



Chi-Square Test of Independence

Death Penalty

		Yes	No	Totals
Race of Victim	White	45	85	130
	Black	14	218	232
	Totals	59	303	362

1. Expected Table under assumption of no relationship between Race of Victim and Death Penalty

	Yes	No
White	21.19	108.81
Black	37.81	194.19

21.19 = (130)(59)/362

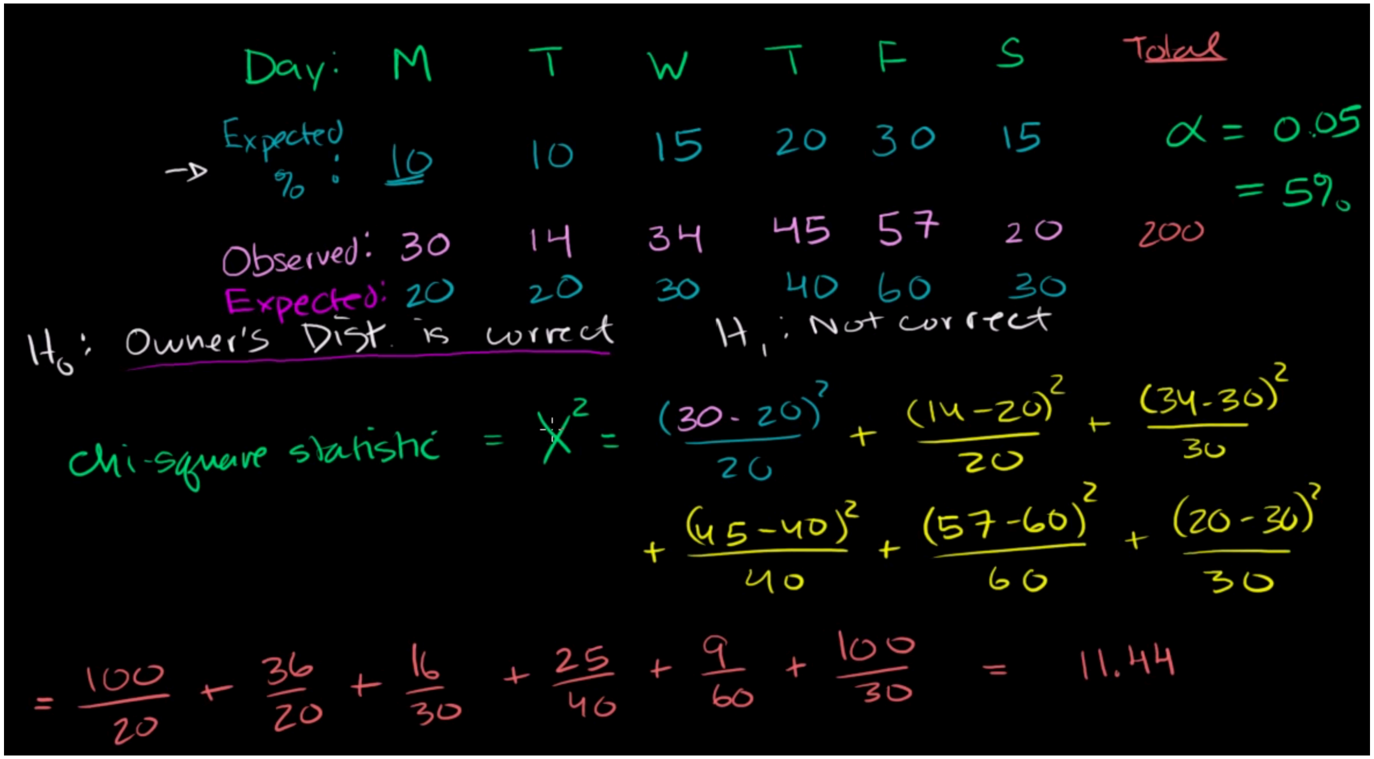
21.19 = (130)(59)/362

2. Compute X^2 test statistic

X^2 = (45-21.19)^2/21.19 + … + (218-194.19)^2/194.19 = 49.89

3. p-value = P(X^2>49.89) = **1.626113e-12** < 0.0001 ➔ Reject null

Chi-Square Goodness of Fit Test



Day:	M	T	W	T	F	S	Total
Expected %:	10	10	15	20	30	15	
Observed:	30	14	34	45	57	20	200
Expected:	20	20	30	40	60	30	

$\alpha = 0.05 = 5\%$

H_0: Owner's Dist. is correct

H_1: Not correct

$$\text{Chi-square statistic} = \chi^2 = \frac{(30-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(34-30)^2}{30} + \frac{(45-40)^2}{40} + \frac{(57-60)^2}{60} + \frac{(20-30)^2}{30}$$

$$= \frac{100}{20} + \frac{36}{20} + \frac{16}{30} + \frac{25}{40} + \frac{9}{60} + \frac{100}{30} = 11.44$$

- Expected is owner's hypothesis about distribution over 6 days
- Observed is some actual data on customer flow
  - We turn this into a X^2 test statistic = 11.44, and compare it to a X^2 distribution with d.f. = 5. ➔ p-value = P(X^2>11.44) = 0.0433 < 0.05 ➔ Reject null distribution

Chi-Square Test of Independence

Death Penalty

		Yes	No	Totals
Race of Victim	White	45	85	130
	Black	14	218	232
	Totals	59	303	362

1. Expected Table under assumption of no relationship between Race of Victim and Death Penalty

	Yes	No
White	21.19	108.81
Black	37.81	194.19

21.19 = (130)(59)/362

21.19 = (130)(59)/362

2. Compute X^2 test statistic

X^2 = (45-21.19)^2/21.19 + … + (218-194.19)^2/194.19 = 49.89

3. p-value = P(X^2>49.89) = **1.626113e-12** < 0.0001 ➔ Reject null

Chi-Square Test

Wait…doesn’t this look a lot like the Two-sample Comparison of Proportions test?

Two-sample Comparison of Proportions

1. State null hypothesis (H_0) and alternative hypothesis (H_A), let α=0.05

   H_0: $p_{white} = p_{black}$

   H_A: $p_{white} \neq p_{black}$

2. Compute appropriate test statistic using collected data.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

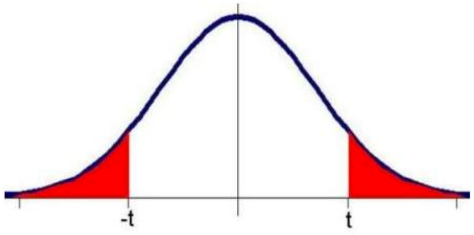
⟹

n_{white}=130, n_{black}=232

$\hat{p}_w = 45/130, \hat{p}_b = 14/232$

z = 7.063152

3. Compute p-value based on test statistic



Since we are doing a two-sided test, we take area to left of z = -7.063152 and right of z = +7.063152

p-value ≈ **1.627673e-12** ⟹ Reject null H_0 in favor of H_A

Experimental vs Observational

- **Experimental**
  - Apply treatments to experimental units (people, animals, land, etc) and observe effect of treatment
  - Can be used to establish causality
  - Example: Randomly assigning hw to students and measuring the performance of the 2 groups

- **Observational**
  - Observe subjects and measure variables of interest without assigning treatments to subjects
  - Can't be used to establish causality
  - Example: Students who did and didn't do their hw and their grades

Confounding factor

- An extraneous attribute that correlates with the dependent variable (performance) and the independent variable (homework or not)

- **What is the confounding factor?**
  - How hard-working the student is
  - More hard-working students might perform better and are more likely to do their hw

Experimental Design

- **Randomization into groups of equal size**
  - Randomly generate number from 0-1
  - If ≤ 0.5, hw, otherwise no hw

- **Assume independent observations**
  - Assume the students don't know if the other students have hw or not
  - Otherwise that knowledge might affect performance

Questions

- What's the general framework of hypothesis testing?
- How to test difference between 2 means?
  - Why does test work even if population distributions from which we sample are quite non-normal?
- How to test difference between 2 proportions?
- Why might one use a Chi-Square test?
- What is the multiple comparisons problem and how do I account for it?
- What is a p-value? Type I error? Type II error?

Questions

- What's the general framework of hypothesis testing?

  H_0,H_A, α → get data, compute test-stat → get p-value
- How to test difference between 2 means? t-test
  - Why does test work even if population distributions from which we sample are quite non-normal? Central Limit Theorem
- How to test difference between 2 proportions? z-test
- Why might one use a Chi-Square test? Goodness of fit, or Test of independence
- What is the multiple comparisons problem and how do I account for it? Bonferroni Correction
- What is a p-value? Type I error? Type II error?
  - p-value is "the probability of observing the data we observed, or more extreme, given the null hypothesis is true"
  - Type I error is P(Reject H_0 | H_0 is true)
  - Type II error is P(Accept H_0 | H_0 is false)

	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)