BROWN
Computer Science

CS1951A: Data Science

Lecture 9: Hypothesis testing

Lorenzo De Stefani

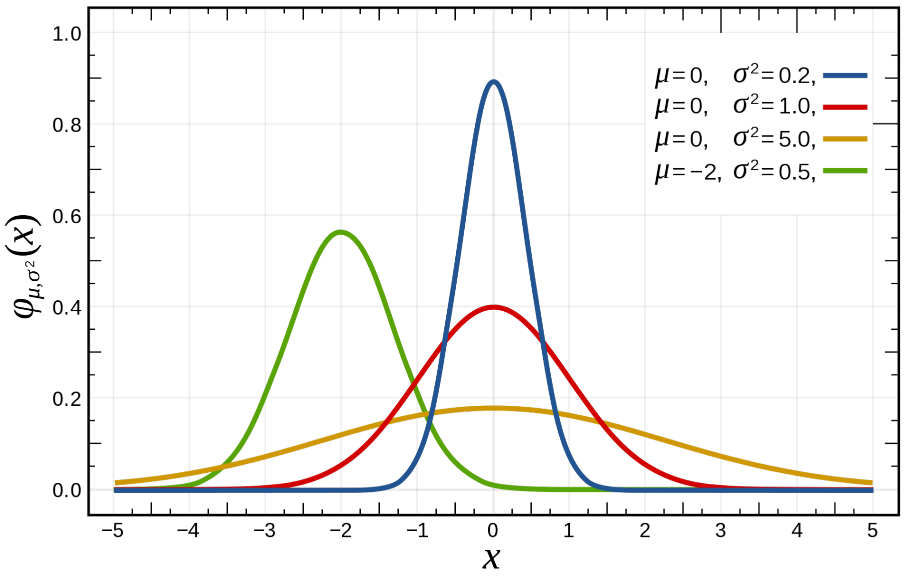
Spring 2022

Outline

- Normal distribution and the central limit theorem
- Testable hypotheses
- A blueprint for the hypothesis testing method
- Testing the fairness of a coin
- P-value and rejection zone
- One side vs two sided hypotheses
- Choosing the correct statistical test
- T-test
- Chi-squared test

Normal or Gaussian Distribution

- Continuous distribution for real-valued random variables of great importance
- Two parameters $X \sim N(\mu, \sigma)$
  - μ expected value
  - σ standard deviation



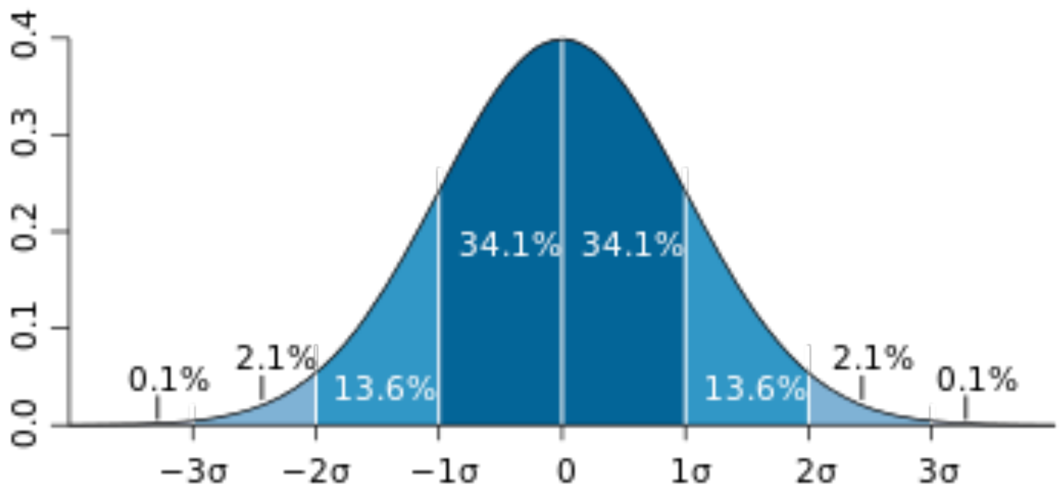
Normal or Gaussian Distribution

- Used to represent many statistical phenomena
  - White noise is normally distributed with mean 0
  - A Normal distribution with $\mu = 0, \sigma = 1$ is called standard normal distribution
- The pmf of a Normal Distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- The cmf quite complex to compute!
  - We generally use tables

Normal or Gaussian Distribution



- The values less than one standard deviation away from the mean account for 68.27% of the set
- Within two standard deviations from the mean account for 95.45%
- Within three standard deviations account for 99.73%.

Law of large numbers: informal statement

- If we repeat the same experiment a large number of times, the average of the outcomes $\overline{X_n}$ (sample average) will converge to the expected value

$$\overline{X_n} = \frac{1}{n}\sum X_i$$

$$\overline{X_n} \rightarrow_{n\rightarrow\infty} \mathrm{E}[X_n] = E[X_i] = \mu$$

- This holds under the assumption that the repetitions X_i are independent and have the same expected value

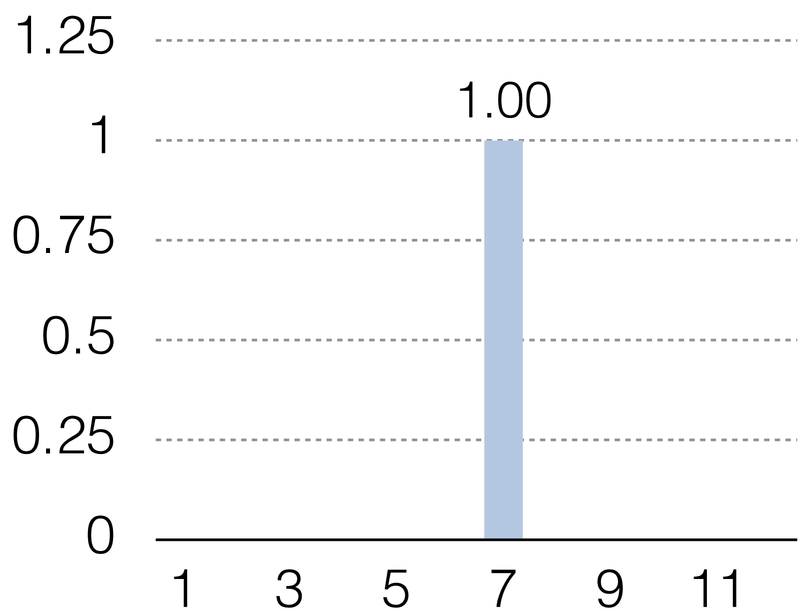
Central limit theorem

The distribution of the sample average $\overline{X_n}$ of n independent and identically distributed samples from a distribution with expected value μ and finite variance σ^2 converges to a normal distribution with expected value μ and variance σ^2/n as $n \to \infty$

- More precisely $\sqrt{n}(\overline{X_n} - \mu)$ approximates $N(0, \sigma^2)$ regardless of the distribution of the samples
- It implies that probabilistic and statistical methods that work for normal distributions can be applied also to many problems involving other types of distributions.

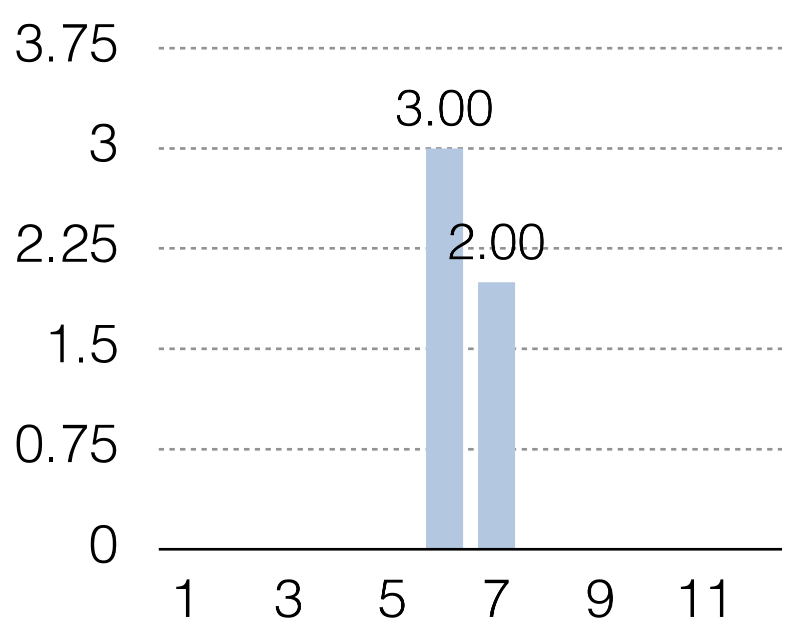
Central Limit Theorem

- Let X be the number of heads obtained when flipping a fair coin 12 times.
- This is binomial random variable with expected value $0.5 \times 12 = 6$
- Repeat the experiment a few times



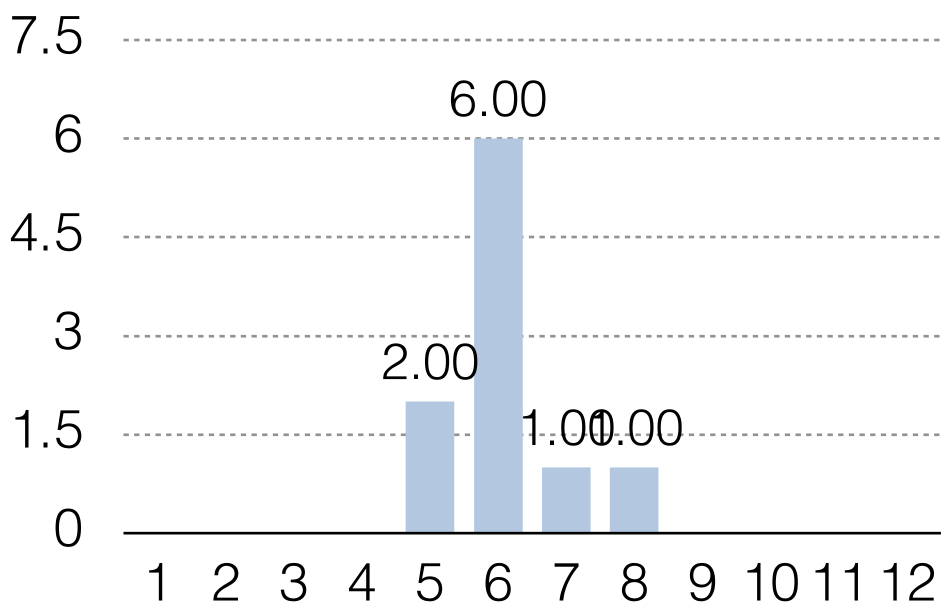
Central Limit Theorem

- Let X be the number of heads obtained when flipping a fair coin 12 times.
- This is binomial random variable with expected value $0.5 \times 12 = 6$
- Repeat the experiment a few times



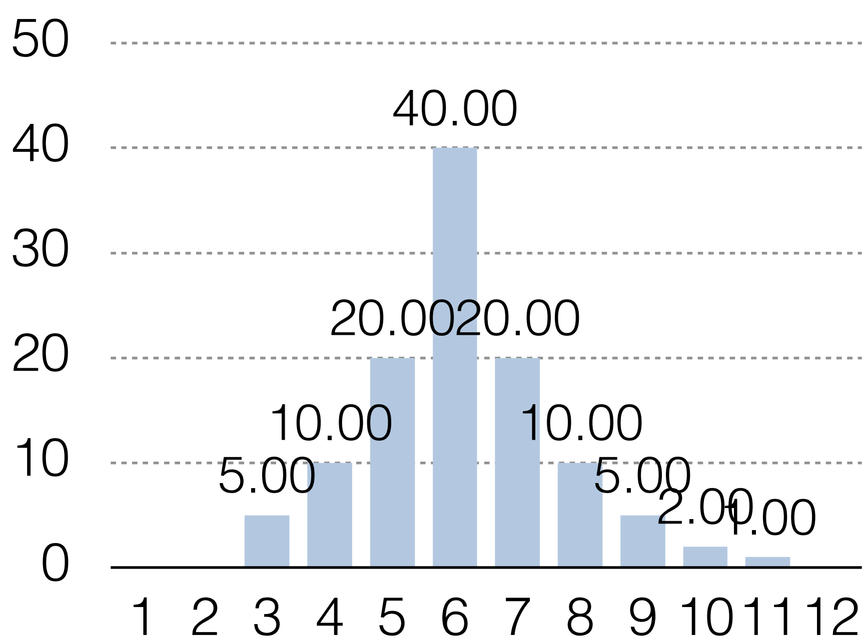
Central Limit Theorem

- Let X be the number of heads obtained when flipping a fair coin 12 times.
- This is binomial random variable with expected value $0.5 \times 12 = 6$



Central Limit Theorem

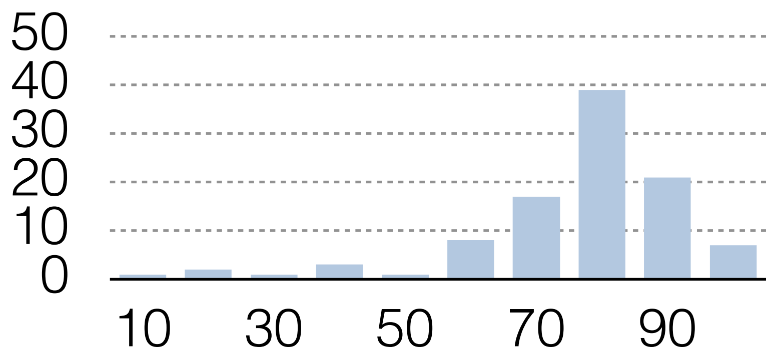
- Let X be the number of heads obtained when flipping a fair coin 12 times.
- This is binomial random variable with expected value $0.5 \times 12 = 6$



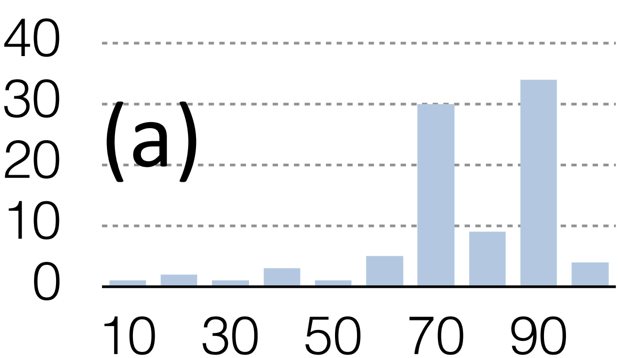
Normal distribution and testing

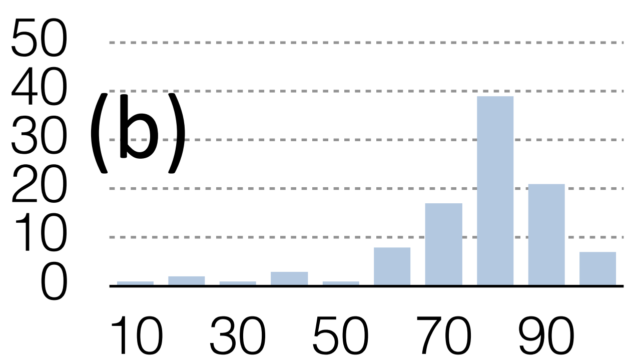
- Testing statistics of interest are often normally distributed
- We can apply statistical methods designed for normal distributions even when underlying distribution is not normal
- We can do so if the statistic converges to the normal distribution as $n \to \infty$

Every year, I compute the mean grade in my class. I never change the material or my methods for evaluating. Over the 439 (☺) years that I have been teaching this class, this has resulted in the below distribution.



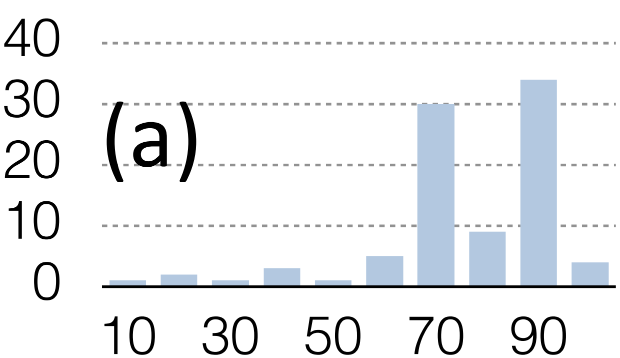
Which of these is mostly like the typical distribution on any given year?

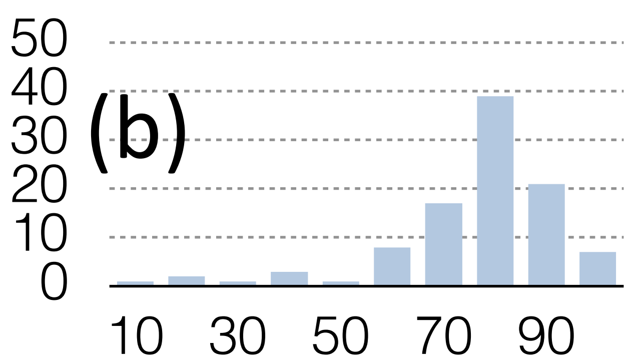




(c) can’t say, could be either

Central Limit Theorem: repeated measures of a mean will be normally distributed. This does not assume the population over which you are taking the mean to be itself normally distributed





(c) can't say, could be either

Null vs. alternative hypothesis

The FDA or “science” needs to decide on a new theory, drug, treatment…

- H_0: The null hypothesis - the current theory, drug, treatment, is as good or better
- H_a: The alternative hypothesis - the new theory, drug, treatment, should replace the old one

Researchers do not know which hypothesis is true. They must make a decision on the basis of evidence presented.

What is a (testable) hypothesis?

- A hypothesis is a claim (assumption) about a population parameter:
  - population mean

    Example: The mean monthly cell phone bill of this city is $\mu = \$42$

  - population proportion

    Example: The proportion of adults in this city with cell phones is $p = .68$

The null hypothesis, H_0

States the assumption (numerical) to be tested

> Example: The average number of TV sets in U.S. Homes is equal to three ($H_0 : \mu = 3$)

Is always about a population parameter, not about a sample statistic

$$H_0 : \mu = 3$$

$$H_0 : \overline{X} = 3$$

The blueprint

- Formulate alternative hypothesis H_a
- Analyze null hypothesis H_0
- Set up experiment
  - Select an appropriate statistical test and a test statistic
  - Come up with a priori theoretical distribution for the test statistic
    - This is is often already given in the definition of the statistical test and H_0
  - Select a threshold $0 \leq \alpha \leq 1$ value for "how surprising" (i.e., how unlikely) under the current assumption H0 the observed data should be in order to decide to reject the null
    - α will denote the level of confidence of the decision
    - If the threshold is used to state the level of confidence which whom we want to decide on rejection
- Acquire data
- Compute the likelihood of observing the test statistic under the null hypothesis
  - p-values!
- Compare the computed value with α and decide if it is possible to reject the null hypothesis

Careful with your terminology!

- Just because we reject a null hypothesis it does not mean we are proving it not to be correct
  - We are merely saying that, given the data, it is unlikely to be correct
  - We can fix the level of confidence of this kind of statement
- Rejecting a null does not imply that the alternative is "correct"
  - Just we cannot exclude it!
  - Avoid the terminology "accepting the alternative"

Example: testing the fairness of a coin

H_1: “this coin is biased”

H_0: “this coin is fair”

Testing procedure

- We flip the coin 20 times independently and with the same distribution
- We count the number of heads called X
  - The “test statistic”
- We compute the probability p of observing a result at least as extreme as X assuming the null hypothesis is correct (“under the null hypothesis”)
  - the p-value
- We set a threshold $0 \leq \alpha \leq 1$ such that if the null hypothesis is rejected if $p \leq \alpha$
  - The desired confidence level

The testing procedure, including the number of samples, type of statistical test and threshold need to be fixed before obtaining the data!!

Level of significance α

- How certain do you want to be?
- Many terminologies: Critical level/control level/critical threshold...
- Example: Significance level of 0.05
  - 5% of the time we will observe higher mean by chance
  - 95% of the time the higher mean will be real
- α bounds the likelihood of making wrong decisions
  - 5% of the time we will reject a correct null by chance

The test

H_0: “this coin is fair”

data

TTHHTTTT
TTTTHTTT
HHTT

test statistic

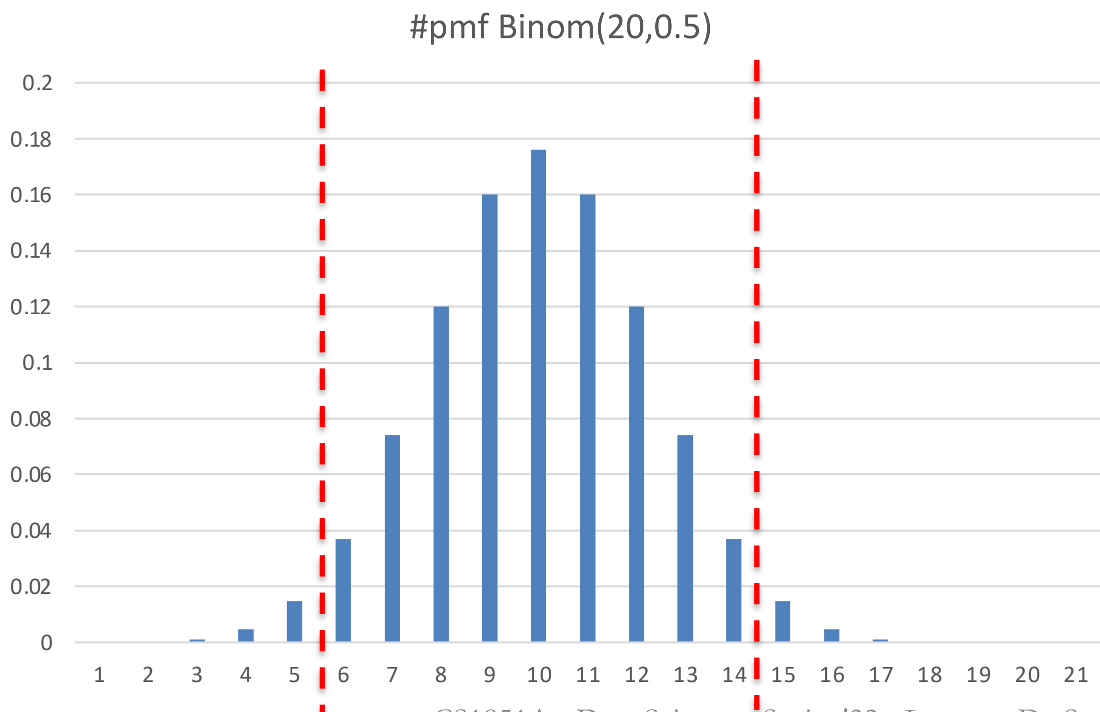
X=5

of “H”

model

$$\langle \Omega, F, P \rangle$$

{H, T} Pr(H)=Pr(T) = 0.5



Probability of a result at least as extreme as X

$$p = 0.03876$$

Set confidence level at α=0.05

$$p \leq \alpha = 0.05$$

H_0 is rejected

The blueprint: example

H_0: "this coin is fair"

data

TTHHTTTT
TTTTHTTT
HHTT

test statistic

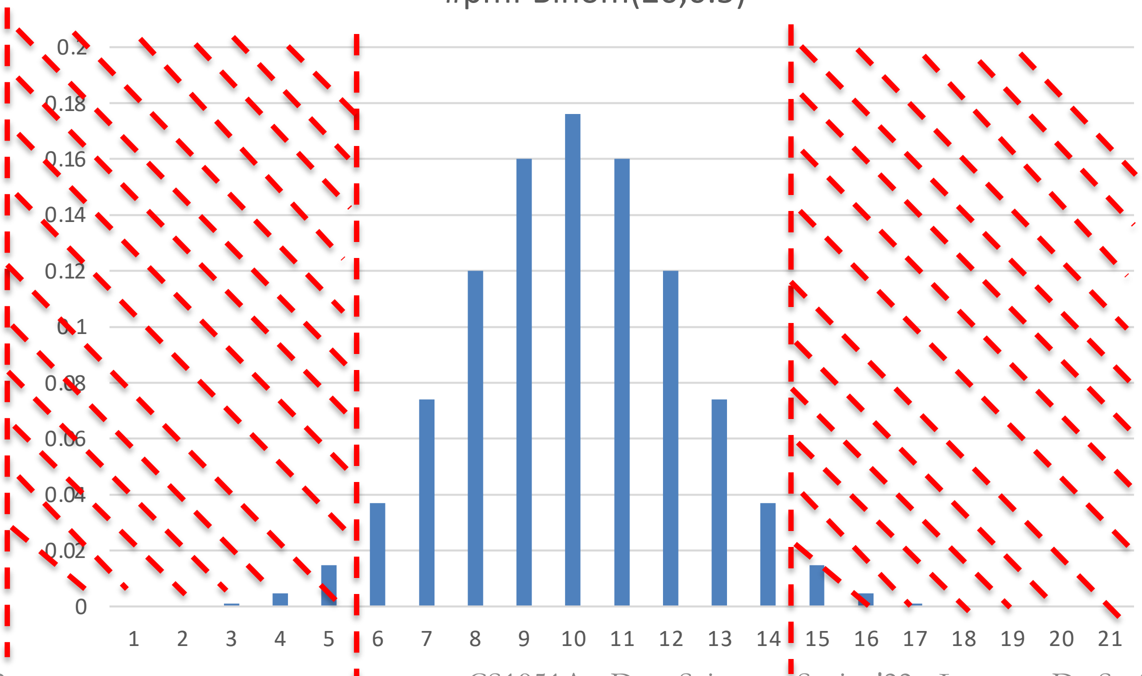
X=5

of "H"

model

$\langle \Omega, F, P \rangle$

{H, T} Pr(H)=Pr(T) = 0.5



We use α to determinate the rejection region

- All the values whose likelihood of being observed is < α
- X is within the rejection region

H_0 is rejected

The blueprint: example

H_0: "this coin is fair"

data

HHTTHTH
TTTTTTHT
HTTTT

test statistic

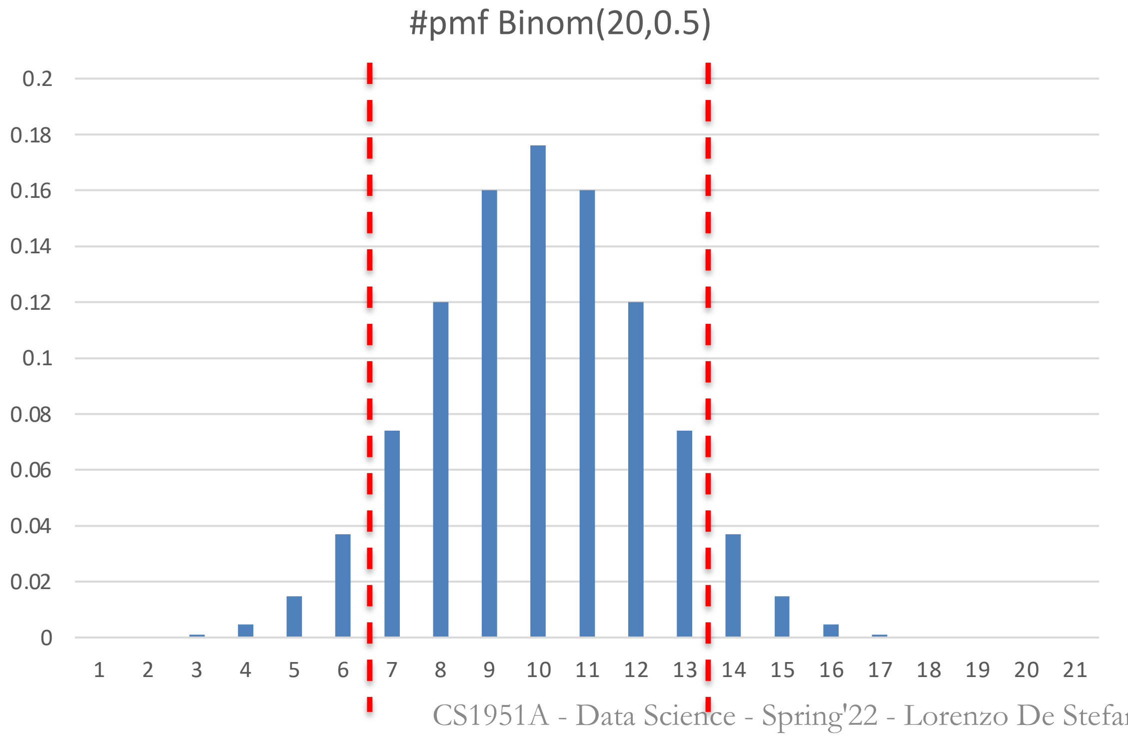
$X=6$

of "H"

$$\langle \Omega, F, P \rangle$$

{H, T} Pr(H)=Pr(T) = 0.5

model



Probability of a result at least as extreme as X

$$p = 0.092$$

Set confidence level at α=0.05

$$p \geq \alpha = 0.05$$

we fail to reject H_0

One tailed and two tailed tests

One-tailed tests

- Based on a unidirectional hypothesis

> Example: The average height of an adult in 2010 is higher than 6 feet

Two-tailed tests

- Based on a bidirectional hypothesis

> Example: The average height of an adult in 2010 different from 6 feet

A slightly different question

H_1: “this coin is biased towards tail”

H_0: “this coin is not biased towards tail”

Testing procedure

- We flip the coin 20 times independently and uniformly at random
- We count the number of heads called X
- We compute the probability p of a result at least as extreme as X under the null hypothesis
  - the p-value
- We set a threshold $0 \leq \alpha \leq 1$ such that if the null hypothesis is rejected if $p \leq a$

The testing procedure, including the number of samples, typo of statistical test and threshold need to be fixed before actually obtaining the data!!

A slightly different question

H_0: "this coin is not biased towards tail"

data

HHTTHTH
TTTTTTHT
HTTTT

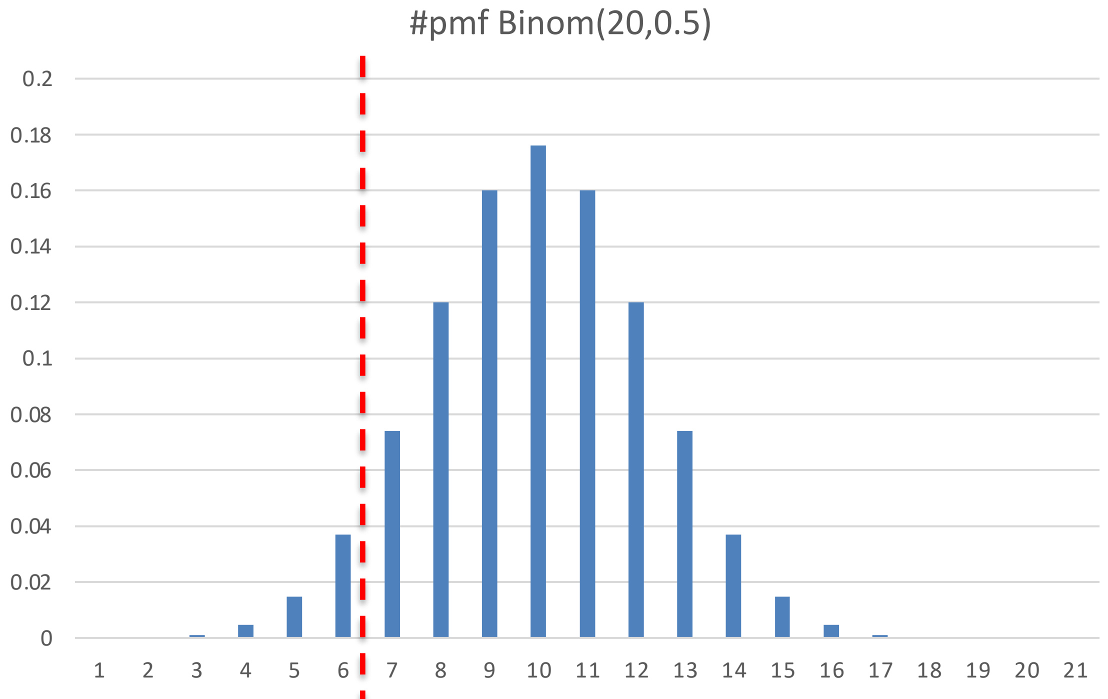
test statistic

$X=6$

of "H"

model

Pr(H)=Pr(T) = 0.5



Unidirectional
Probability of a result at least as extreme as X

$$p = 0.046$$

Set confidence level at α=0.05

$$p \geq \alpha = 0.05$$

we reject H_0

A slightly different question

H_0: "this coin is not biased towards tail"

data

HHTTHTH
TTTTTTHT
HTTTT

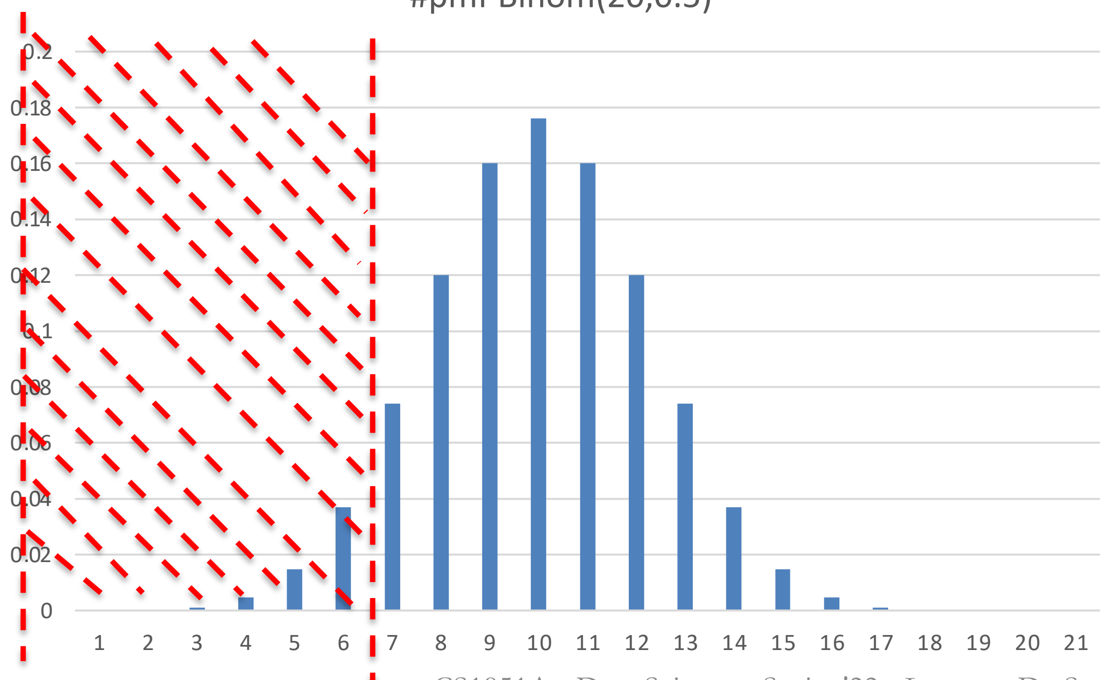
test statistic

X=6

of "H"

model

Pr(H)=Pr(T) = 0.5



Set rejection region according to unidirectional test

X is in the rejection region

we reject H_0

One-tailed test

p-value

- p-value: Probability of obtaining a test statistic more extreme (≤ or ≥) than the observed sample value given H_0 is true
  - Also called observed level of significance
  - Function of the data takes values in [0,1]

Uniform random variable

- A discrete random variable X which takes values in $\{x_1, x_2, \dots, x_m\}$ is uniform if

$$\forall x_i \in \{x_1, x_2, \dots, x_m\}, \Pr(\mathrm{X} = \mathrm{x_i}) = \frac{1}{|\{x_1, x_2, \dots, x_m\}|}$$

  - $\mathrm{E}[\mathrm{X}] = \frac{\sum_{i=1}^{m} x_m}{|\{x_1, x_2, \dots, x_m\}|}$

- A continuous random variable X in the interval [a,b] , denoted as $U(a, b)$, is such that given $c \in [a, b]$

$$\Pr(X \leq c) = \frac{c - a}{b - a}$$

  - $\mathrm{E}[\mathrm{X}] = \frac{b-a}{2}$

Another view of p-value

- As $n \to \infty$, the p-value p is a random variable whose pmf is uniform in [0,1] if the null-hypothesis is correct
- Assume we set $\alpha = 0.05$:
  - $\Pr(p \leq 0.05) = 0.05!$

One sample vs two samples test

- One sample tests:
  - We have observations (samples) from one population, we want to compare them with a fixed model or distribution
  - E.g., this distribution as mean μ
- Two samples tests:
  - We have observations (samples) from two populations
  - We want to compare statistical properties of the two populations through the observations
  - E.g., these two distributions are the same, they have the same average, ...

Many important test have both versions

Select your test

- Testing is a bit like finding the right recipe based on these ingredients:
  - Type of hypothesis
  - Data type
  - Sample size
  - Variance known? Variance of several groups equal?
- Good news: Plenty of tables available, e.g.,
  - http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm (with examples in R, SAS, Stata, SPSS)
  - http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf

Select your test

Population

Example of a table of tests

Summary Table for Statistical Techniques

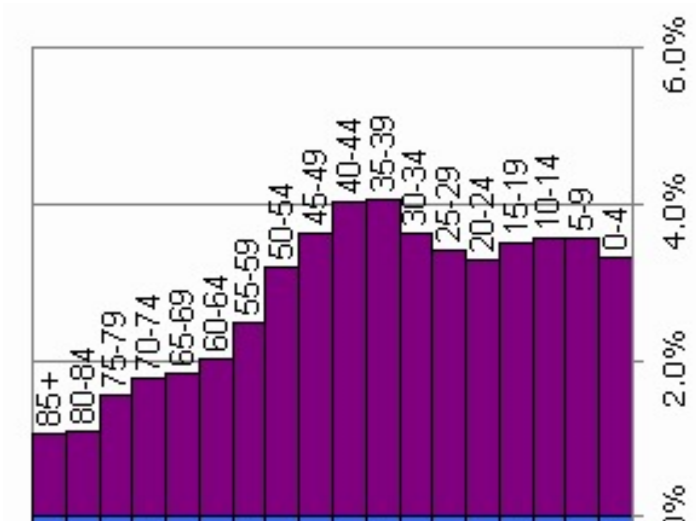
	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
1	Estimating a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• What is the average weight of adults? • What is the average cholesterol level of adult females?	1-sample t-interval $\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately normal or • have a large sample size ($n \geq 30$)
2	Test about a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• Is the average GPA of juniors at Penn State higher than 3.0? • Is the average Winter temperature in State College less than 42° F?	$H_o: \mu = \mu_o$ $H_a: \mu \neq \mu_o$ or $H_a: \mu > \mu_o$ or $H_a: \mu < \mu_o$ The one sample t test: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately normal or • have a large sample size ($n \geq 30$)
3	Estimating a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• What is the proportion of males in the world? • What is the proportion of students that smoke?	1-proportion Z-interval $\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	Stat >Basic statistics >1-sample proportion	• have at least 5 in each category
4	Test about a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• Is the proportion of females different from 0.5? • Is the proportion of students who fail Stat 500 less than 0.1?	$H_o: \pi = \pi_o$ $H_a: \pi \neq \pi_o$ or $H_a: \pi > \pi_o$ or $H_a: \pi < \pi_o$ The one proportion Z-test: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Stat >Basic statistics >1-sample proportion	• $n \pi_o \geq 5$ and $n(1-\pi_o) \geq 5$

http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf

Some tests you are likely to use

- **t-test**: difference of means; is the average value of some feature different between two populations
  - e.g., Are men taller than women? Are blue states more populated than red states? Do CS students work harder than other majors?
- **chi-squared X^2-test**: difference in frequencies of a categorical variable; is the distribution of some feature uniform across groups
  - Used to compare distributions of discrete random variables: for continuous ones is better to use Kolmogorov-Smirnoff test
  - e.g. Do neighborhoods differ in terms of music preferences? Do college majors differ in terms of sociodemographic features?

t-test: example on population means



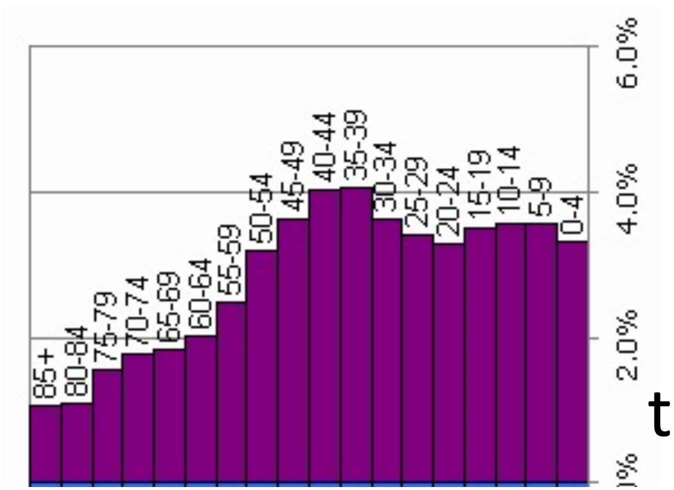
Distribution of ages in the US

H_a= Mean age is not 35

H_0 = Mean age is 35

http://www.censusscope.org/us/chart_age.html

t-test: example on population means



sample average

expectation according to the null

test statistic $z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$

\# of points

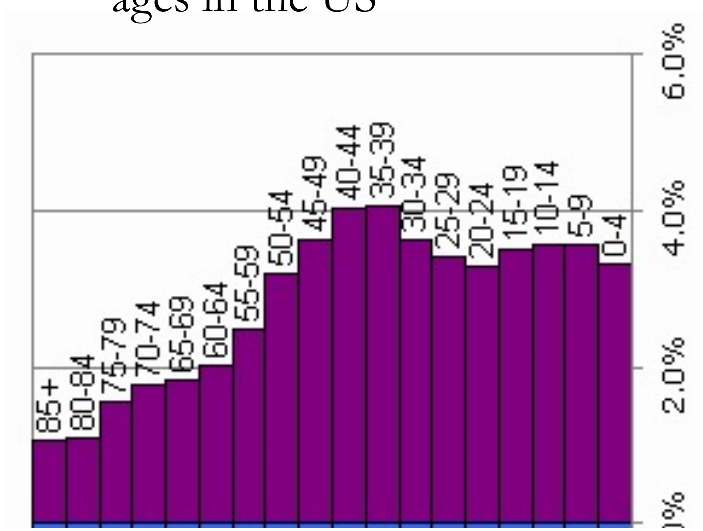
variance

We assume data points are independent samples with (by H_0) same expectation $\mu_0 = 35$

H_0 = Mean age is 35

t-test: example on population means

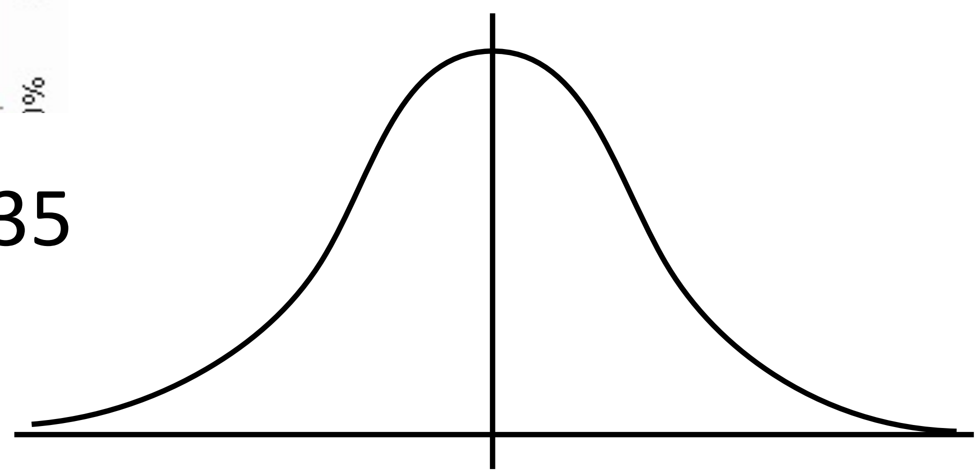
Distribution of
ages in the US



H_0 = Mean age is 35

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

z converges to $N(0,1)$
as $n \to \infty$

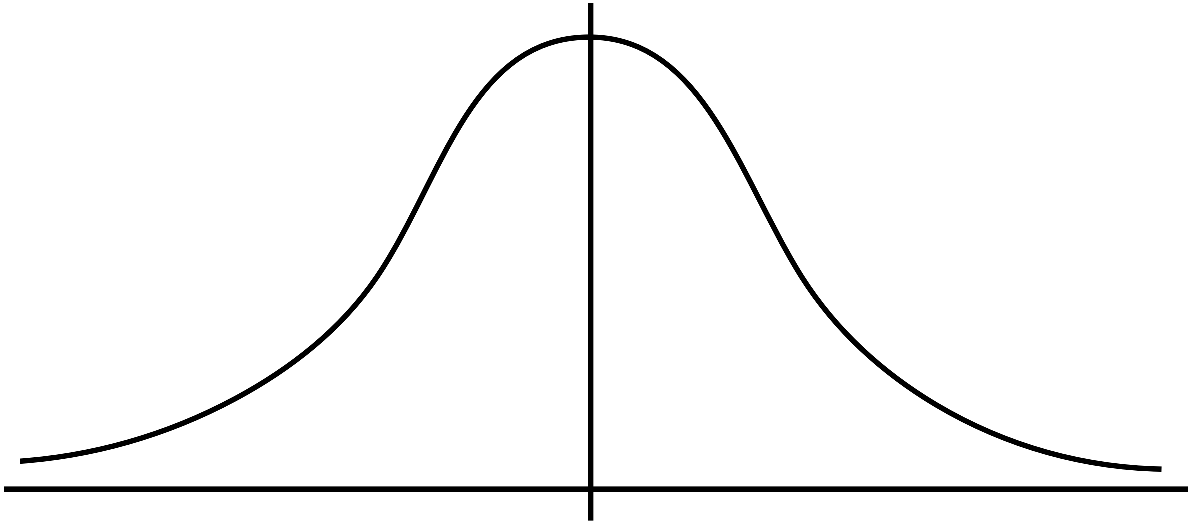


Why can we use a normally-distributed test statistic to evaluate mean age of a population?

a) Because ages are normally distributed
b) Because the test statistic is a random variable
c) Because of the law of large numbers
d) Because of the central limit theorem
e) The limit does not exist!

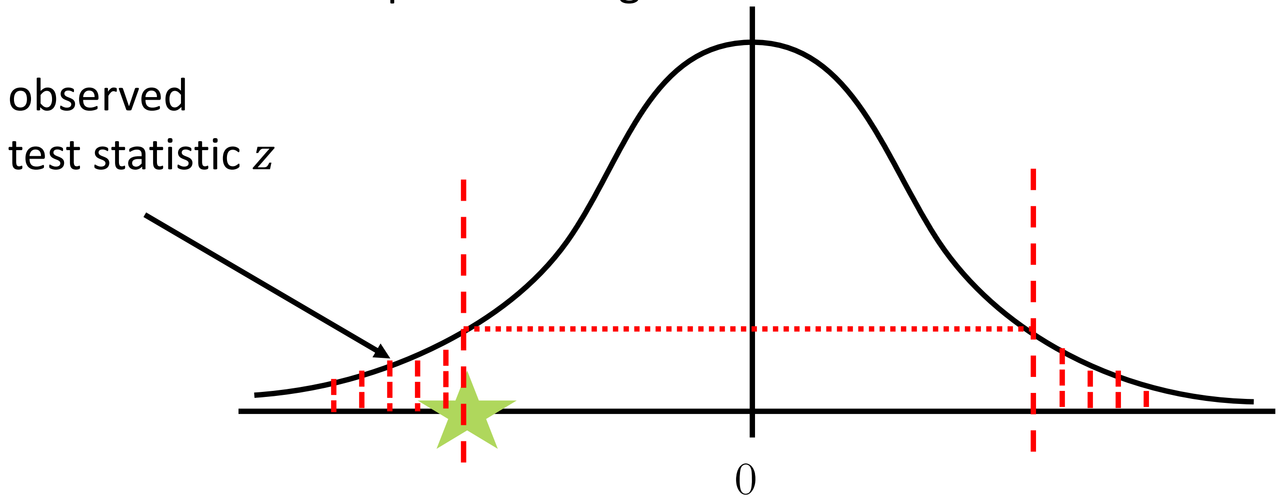
t-test: example on population means

- $z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\bar{X} - \mu_0}{s}$ distance from mean in std units
- the pmf of z is $\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$



t-test: example on population means

- z is the test statistic not the p-value
- the p-value is the likelihood of observing a phenomenon at least as extreme as the one reported by the test statistic assuming null hypothesis is correct
  - In the example assuming we have a two-tailed test



- The red shaded area is the p-value
- How do we compute it? We use tables ☺
- p-value is then compared with threshold α
  - if $p - val \leq \alpha$ reject H_0
  - otherwise fail to reject H_0

t Table

degrees of freedom

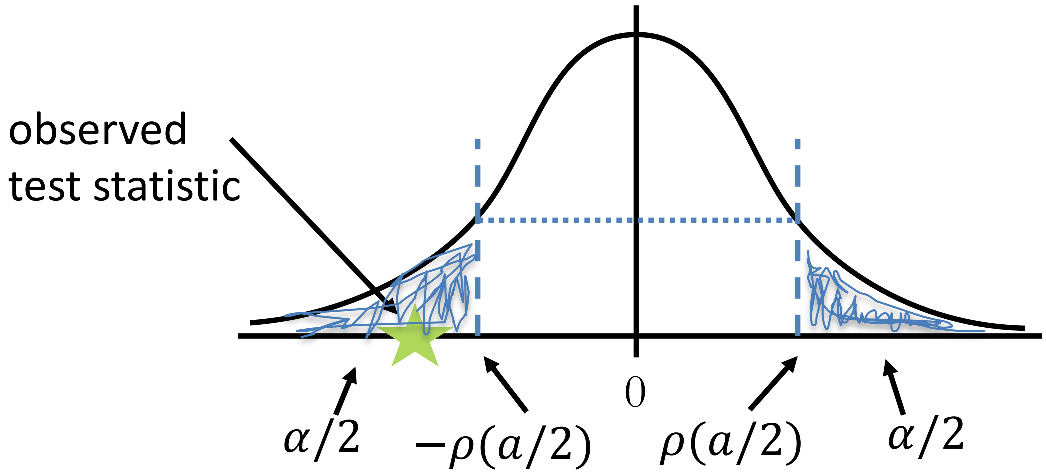
$n - 1$

p-vaue

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

t-test: example on population means

- alternatively given the threshold α, we can compute the boundaries for the rejection zone



- Two tailed test α area of the two tails
- $\rho\left(\frac{a}{2}\right)$ computed using the corresponding table
- We use use $\rho(a/2)$ as criterion
  - $|z| \geq \rho(a/2)$ reject H_0
  - otherwise fail to reject

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

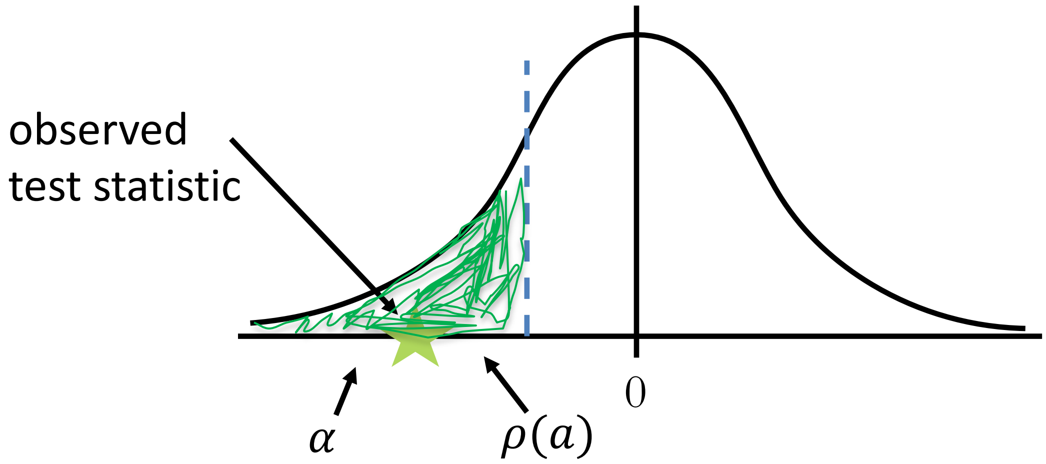
degrees of freedom

$n - 1$

α

t-test – one tailed test

- H_a = "the average age is less than 35 "
- computation of test statistic does not change



- The change impacts evaluation of the p-value or evaluation of $\rho(a)$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

degrees of freedom

$n - 1$

α

What looks suspicious?

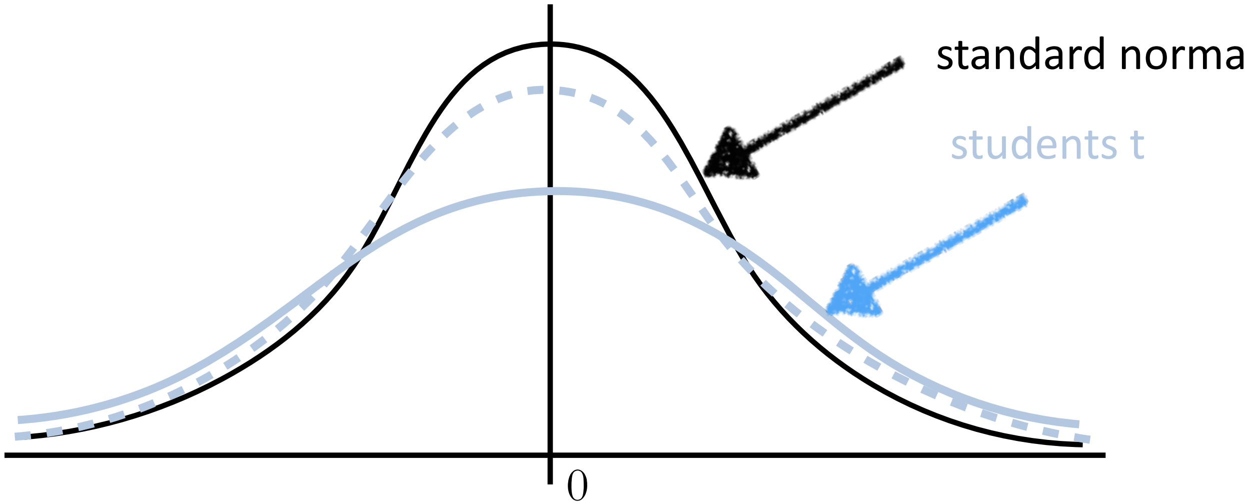
$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

- Generally we do not know the value σ
- We need to replace it with something that can be evaluated from the data

Using empirical standard deviation

$$z = \sqrt{n}\frac{\bar{X} - \mu_0}{s}$$

$$s = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

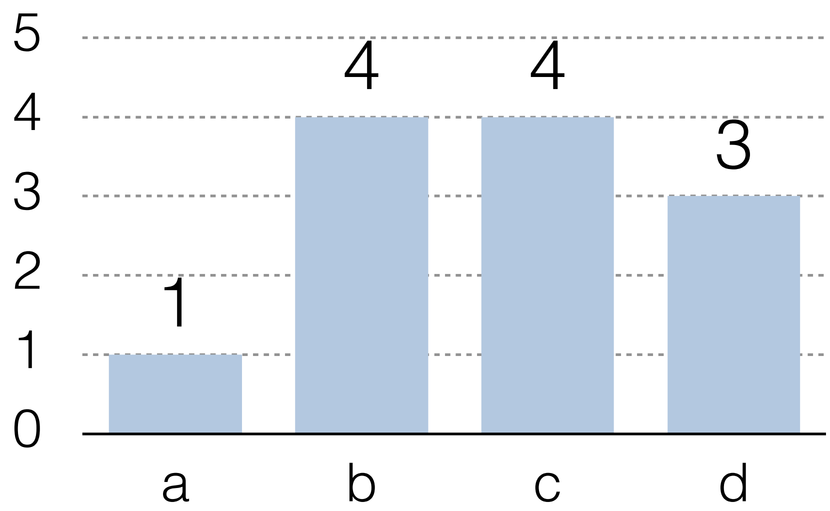


- Students' t distribution converges to the standard normal ad n increases
  - Central limit theorem!
- In general, good convergence for $n \geq 30$

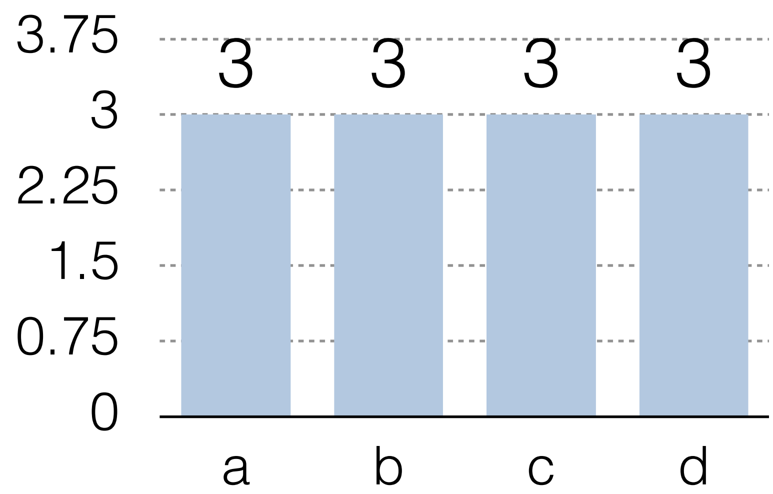
Some tests you are likely to use

- **t-test**: difference of means; is the average value of some feature different between two populations
  - e.g. are men taller than women, are blue states more populated than red states, do CS work harder than other majors
- **chi-squared test**: difference in frequencies of a categorical variable; is the distribution of some feature uniform across groups
  - e.g. do neighborhoods differ in terms of music preferences features; do college majors differ in terms of sociodemographic features
  - Used to compare distributions of discrete random variables: for continuous ones is better to use Kolmogorov-Smirnoff test

Are the answers to driving licence test random?



Observed data
$n = 12$



Expected values assuming H_0 correct for $n = 12$

H_0= all answers are equally likely

I can reframe the question as: is the observed discrete distribution the same as one which uniform over the same values (i.e., the domain)

The X^2 test statistic

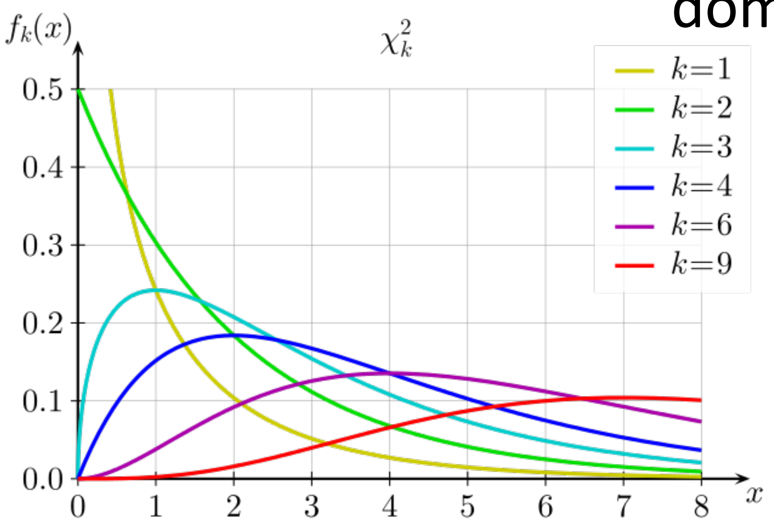
values in the domain of the distribution

occurrences of the i-th value in the domain

$$X^2 \text{ test statistic } Z = \sum_{i=1}^{k} \frac{(c_i - E[c_i])^2}{E[c_i]}$$

Expected # of occurrences of the i-th value in the domain under H_0

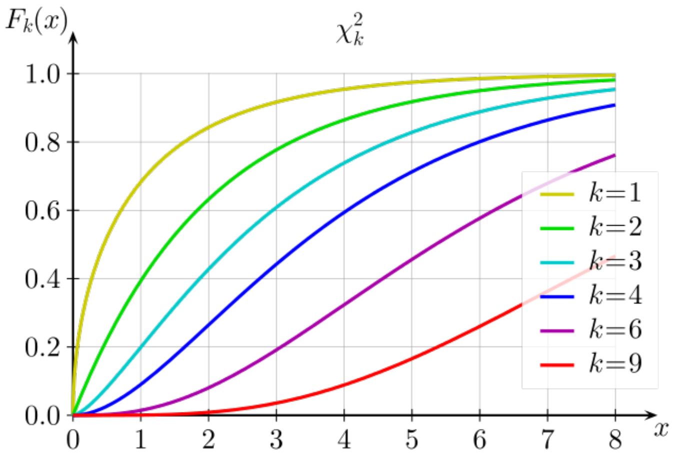
Pearson showed that z's distribution converges to X^2 distribution as $n \to \infty$



pmf X^2 dsitribution

Chi Squared Test

- Compute test statistic $z = \sum_{i=1}^{k} \frac{(c_i - E[c_i])^2}{E[c_i]}$
- We compute the p-value using the cdf of the Chi-squared distribution



- We use tables!

Chi Squared Test

Degrees of freedom are the values that the discrete distribution being observed can assume +1

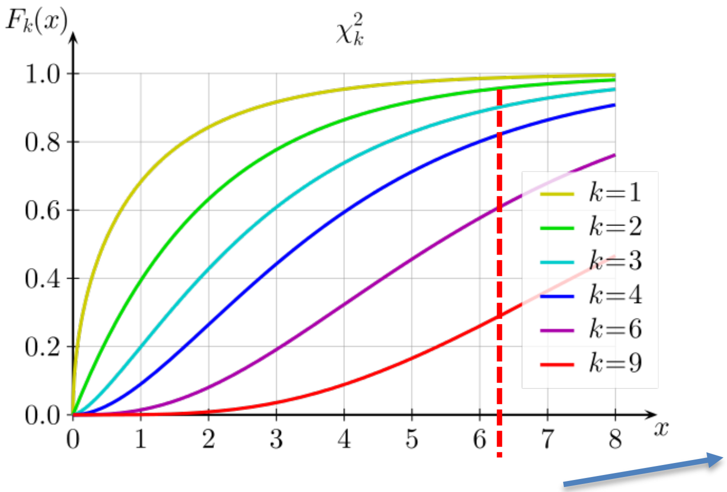
Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of Freedom	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi Squared Test

- Compute test statistic $z = \sum_{i=1}^{k} \frac{(c_i - E[c_i])^2}{E[c_i]}$
- We compute the p-value using the cdf of the Chi-squared distribution



for $\alpha = 0.2$
threshold $\phi(\alpha)$
rejection region to the right

- We use tables!
- For a given significance level α we have a corresponding rejection region
- if $z \geq \phi(\alpha)$ reject null hypothesis with confidence α

Add this to your calendar

Featured Event

Causality in Data Science →

Featuring

2021 Nobel Prize in Economics Winner Guido Imbens '91 Ph.D.

The Applied Econometrics Professor and Professor of Economics Graduate School of Business, Stanford University

Wednesday, March 2, 2022 | 5 p.m.

Salomon Center for Teaching, De Ciccio Auditorium, Room 101
79 Waterman Street

A conversation moderated by Provost Richard M. Locke, Schreiber Family Professor of Political Science and International and Public Affairs, will follow the talk.

Speaker Biography

Register Now