Review of Statistical Inference

Zack W Almquist (University of Washington)

2023-01-07

Contents

Preliminaries

Definition *Data*: Data (in this class) will be a vector or matrix of values. E.g.,

	1940	1945	1950	1955	1960
Food and Tobacco	22.20	44.50	59.60	73.20	86.80
Household Operation	10.50	15.50	29.00	36.50	46.20
Medical and Health	3.53	5.76	9.71	14.00	21.10
Personal Care	1.04	1.98	2.45	3.40	5.40
Private Education	0.34	0.97	1.80	2.60	3.64

Definition *Statistic*: A statistic $t(Y)$ is any function of the data. E.g.,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Definition *Descriptive data analysis*: A representation of the main features of a dataset via a set of statistics $t_1(Y), \ldots, t_k(Y)$. E.g.,

- Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Proportion: $\hat{p} = \frac{Part}{Total} = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in G\}$
- Median: Half the population is above and half below VALUE
- Mode: Most common VALUE
- Variance: $\hat{Var}(X) = \frac{1}{n-1}\sum i = 1^n (X_i - \bar{X})^2$
- Standard Deviation: $\hat{SD}(X) = \sqrt{\hat{Var}(X)}$
- Standard Error: $SE(\bar{X}) = \sqrt{\hat{Var}(X)/n}$

Definition *Test statistic*: A *test statisti* is a standardized value that is calculated from sample data durning a hypothesis test.

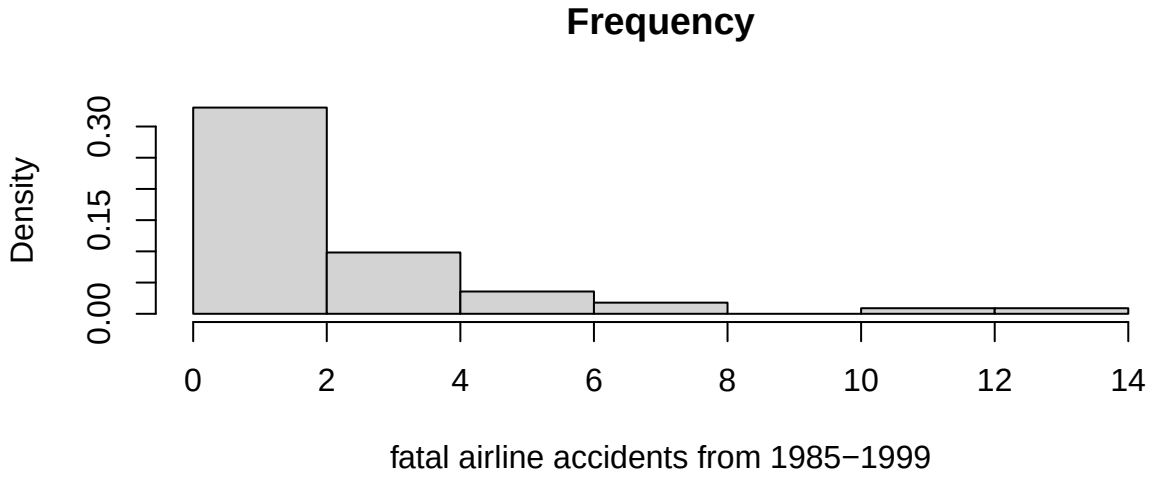
Definition *Statistical Hypothesis test*: Evaluates two mutually exclusive statements about a population to determine which statement is best supported by the (sampled) data.

Definition *Distribution*: For this class distribution is the histogram or normalized frequency counts of our data. For example: Say we have the dataset of airlines by fatal accidents from 1985-199. (Below is the first 5 rows)

	airline	fatal_accidents_85_99
1	Aer Lingus	0
2	Aeroflot	14
3	Aerolineas Argentinas	0
4	Aeromexico	1
5	Air Canada	0

Then the distribution of fatal airline accidents from 1985-1999 would be the histogram of this data,

With the Frequency histogram being:



and the normalized (disribution) histogram being:

What are we doing and why?

An introduction to statistical inference is really an introductuction to *statistical decision* making. To make a decision we need to be able to evaluate the *meaning* of our descriptive statistics. Typically this takes two majore forms: (1) comparison of two groups or (2) comparison against an assumed baseline.

Hypothesis testing algorithem

1. Question that can be answered with data. (E.g., is gender discrimination occuring in corperation the Tech sector?)
2. State **null** (usually 0 or status quo) and state **alternative** hypothesis (depends on question).
   - Make sure and also write down the desriptive statistic or comparison statistics of interest.
3. Perform a *statistical test*.
4. Evaluate evidence.

Writing a Null versus Alternative Hypothesis

- Language for null hypothesis:
  - H_0: The null hypothesis of BLANK is that BLANK is VALUE or 0.
- Language for the alternative hypothesis:
  - H_A: The alternative hypothesis is that BLANK is $>$ VALUE, $<$ VALUE or $\neq$ VALUE.

BLANK is your statistics (e.g., mean value of X).

Test statistic and evaluation of the Hypothesis

So far in class we have two *Null* hypothesis we can ask:

- H_0: Label BLANK does not matter.
- H_0: The STATISTIC of BLANK is VALUE or 0.

with

- H_A: Label BLANK is related to OUTCOME BLANK.
- H_A: The STATISTIC of BLANK is $>$ VALUE, $<$ VALUE or $\neq$ VALUE.

Simulation/Permutation Test for Labeled data Example 1

The Problem (Taken From Fivethirtyeight). Does education level affect wether one cares about the oxford comma?

	No	Yes	Sum
Associate degree or less	248	158	406
Bachelor degree or higher	326	294	620
Sum	574	452	1026

Hypothesis (one tail):

- H_0: Label EDUCATION does not matter. ($p_{BDH} - p_{ADL} = 0$)
- H_A: Label EDUCATION matters and we expect $p_{BDH} > p_{ADL}$.
- H_A: Label EDUCATION matters and we expect $p_{BDH} < p_{ADL}$.

Test Statistic is:

$$\hat{p}_{BDH} - \hat{p}_{ADL} = \frac{294}{620} - \frac{158}{406} = 0.085$$

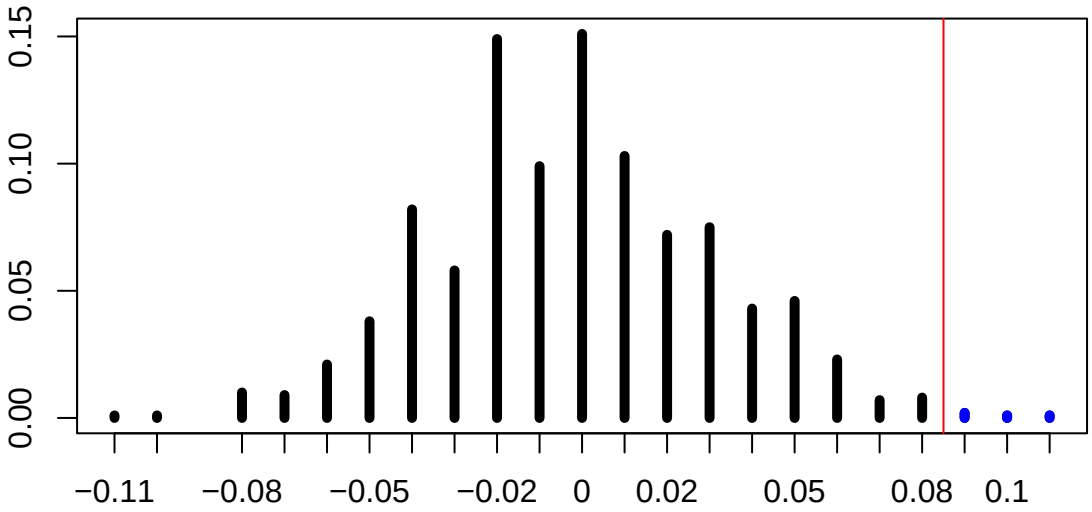
Simulation/Permutation Procedure

- We assume the (NULL) labels have no meaning, so if we randomly re-assign the label to the outcome data we can acquire our NULL distribution.
  - If the observed test statistics is in the center of this distribution then we have strong evidence of the NULL hypothesis
  - If the observed test statistics is in the tails then we have evidence of the alternative hypothesis.

Simulated Test (10 Simulations)

```
 [1]  0.033  0.033 -0.012 -0.040  0.037  0.017 -0.048 -0.048 -0.008 -0.032
```

Resulting plot (1000 simulations):



Evaluation

- H_A: $p_{BDH} > p_{ADL}$

Sim.Test
$0.033 > 0.085 = 0$
$0.033 > 0.085 = 0$
$-0.012 > 0.085 = 0$
$-0.04 > 0.085 = 0$
$0.037 > 0.085 = 0$

Sim.Test
$0.017 > 0.085 = 0$
$-0.048 > 0.085 = 0$
$-0.048 > 0.085 = 0$
$-0.008 > 0.085 = 0$
$-0.032 > 0.085 = 0$

Evaluation We reject the null. We have strong evidence that H_A is true as the observed value is always higher than the randomly labeled simulation.

- H_A: $p_{BDH} < p_{ADL}$

Sim.Test
$0.033 < 0.085 = 1$
$0.033 < 0.085 = 1$
$-0.012 < 0.085 = 1$
$-0.04 < 0.085 = 1$
$0.037 < 0.085 = 1$
$0.017 < 0.085 = 1$
$-0.048 < 0.085 = 1$
$-0.048 < 0.085 = 1$
$-0.008 < 0.085 = 1$
$-0.032 < 0.085 = 1$

Evaluation We accept the null. We have no evidence that H_A is true as the observed value is always higher than the randomly labeled simulation.

Hypothesis (two tail):

- H_A: Label EDUCATION matters and we expect $p_{BDH} \neq p_{ADL}$.

Sim.Test
$0.033 > \mid 0.085 \mid = 0$
$0.033 > \mid 0.085 \mid = 0$
$-0.012 > \mid 0.085 \mid = 0$
$-0.04 > \mid 0.085 \mid = 0$
$0.037 > \mid 0.085 \mid = 0$
$0.017 > \mid 0.085 \mid = 0$
$-0.048 > \mid 0.085 \mid = 0$
$-0.048 > \mid 0.085 \mid = 0$
$-0.008 > \mid 0.085 \mid = 0$
$-0.032 > \mid 0.085 \mid = 0$

Where $|\cdot|$ is the absolute value function (i.e., $|-5| = 5$).

Evaluation

We have *strong* evidence that $p_{BDH} \neq p_{ADL}$ is true as the absolute value of observed value is always higher than the randomly labeled simulation.

Statistical Decision Making

Now say we want to formalize our **evaluation** of our hypothesis and provide a since of how confident we are in our conclusion.

Statistical Significance (α-level and P-Value)

We say that we have *Statistica Significance* (i.e., we have evidence our H_A is correct) if our **P-Value** is less than some α, and that the test is **NOT Statistically Significant** if the **P-Value** is greater than α.

Calculating P-Value (Simulation/Permutation Test)

To calculate the **P-Value** of our permutation test we simply sum uo the number of times our **Test Statistic** is greater than the *simulated* values.

Example (One-tail)

H_A: $p_{BDH} > p_{ADL}$

Sim.Test
$0.033 > 0.085 = 0$
$0.033 > 0.085 = 0$
$-0.012 > 0.085 = 0$
$-0.04 > 0.085 = 0$
$0.037 > 0.085 = 0$
$0.017 > 0.085 = 0$
$-0.048 > 0.085 = 0$
$-0.048 > 0.085 = 0$
$-0.008 > 0.085 = 0$
$-0.032 > 0.085 = 0$

$$P - Value = \frac{0+0+0+0+0+0+0+0+0+0}{10} = 0$$

Conclusion: Statistically Significant

Example (Two-tail)

H_A: $p_{BDH} \neq p_{ADL}$.

Sim.Test
$0.033 > \mid 0.085 \mid = 0$
$0.033 > \mid 0.085 \mid = 0$
$-0.012 > \mid 0.085 \mid = 0$
$-0.04 > \mid 0.085 \mid = 0$
$0.037 > \mid 0.085 \mid = 0$
$0.017 > \mid 0.085 \mid = 0$
$-0.048 > \mid 0.085 \mid = 0$
$-0.048 > \mid 0.085 \mid = 0$
$-0.008 > \mid 0.085 \mid = 0$
$-0.032 > \mid 0.085 \mid = 0$

$$P - Value_{\text{one tail}} = \frac{0+0+0+0+0+0+0+0+0+0}{10} = 0$$

$$P - Value_{\text{two tail}} = 2 \times P - Value_{\text{one tail}} = 0$$

Conclusion: Statistically Significant

Selecting α-level

Traditionally $\alpha = 0.05$. If α is NOT SPECIFIED then $\alpha = 0.05$. Otherwise α can vary – we will cover why and when we cover **Type 1** and **Type 2** errors.

Example (One-tail)

H_A: $p_{BDH} > p_{ADL}$

- $P - Value_{\text{one tail}} = 0 < \alpha = 0.05$. *Statistically Significant*

Example (Two-tail)

H_A: $p_{BDH} \neq p_{ADL}$.

- $P - Value_{\text{two tail}} = 0 < \alpha = 0.05$. *Statistically Significant*

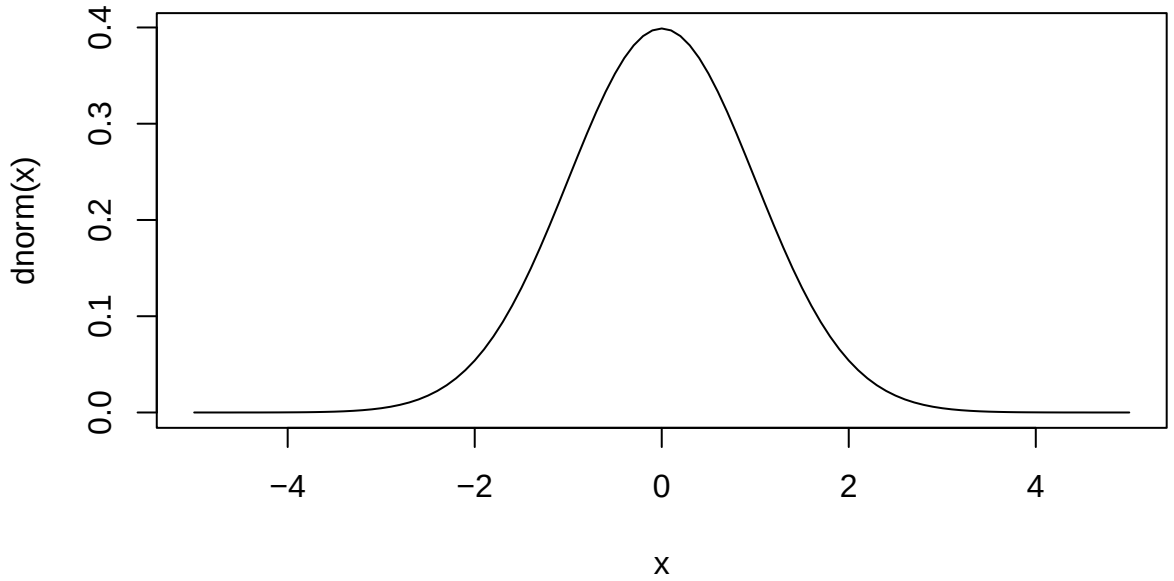
Normal Approximation

What if we want to test more complex hypothesis than does LABEL matter?

- We cand do this! We will use the Normal Distribution to evaluate our descriptive statistics
- Second Question: What is the Normal Distribution?

The Normal Distribution

The Normal Distribution (or Gaussian Distribution) is the classic *unimodal*, *symmetric* distribution that typically looks as follows:



- The Normal Distribution is defined by its mean (μ) and its standard diviation (SD) (σ) [or by its variance σ^2]

- Short hand $N(\mu, \sigma)$

Properties of Normal Distributions

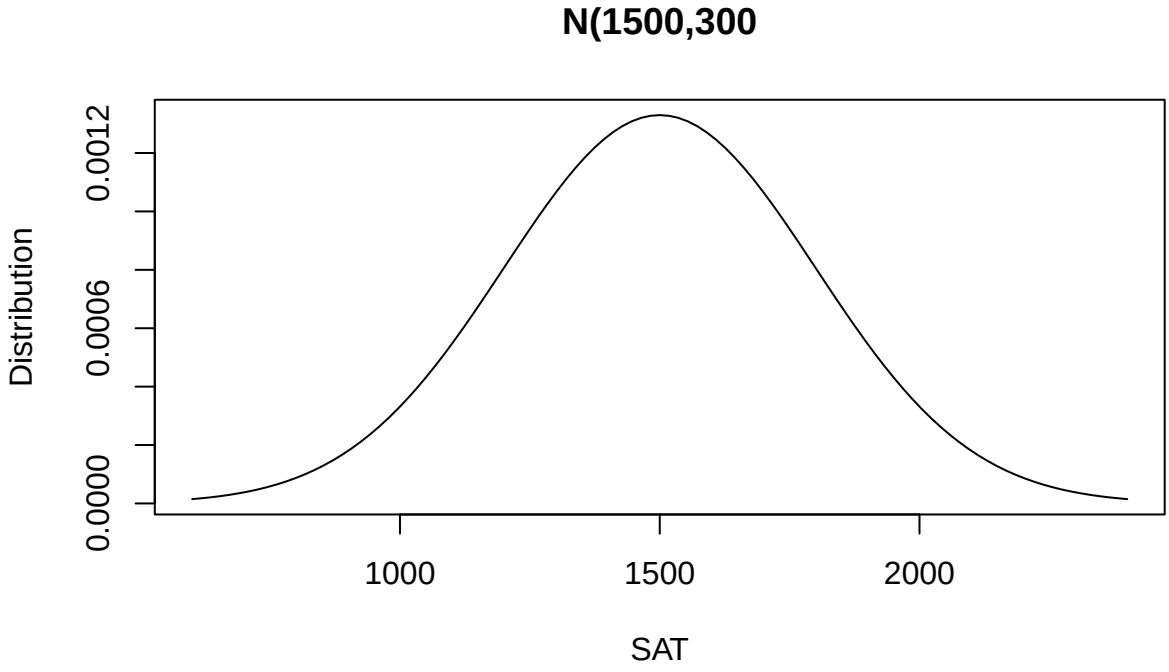
- We can Transform normal distributions and perserve key identities (i.e., if $x_1 < x_2$ and we transform the data $x_1^* < x_2^*$, etc)
- What does this mean in practice?
  - If we transform out observed mean (X) into a known distribution (e.g., Z) we can readily perform our hypothesis tests!

Exampe: SAT and ACT Scores

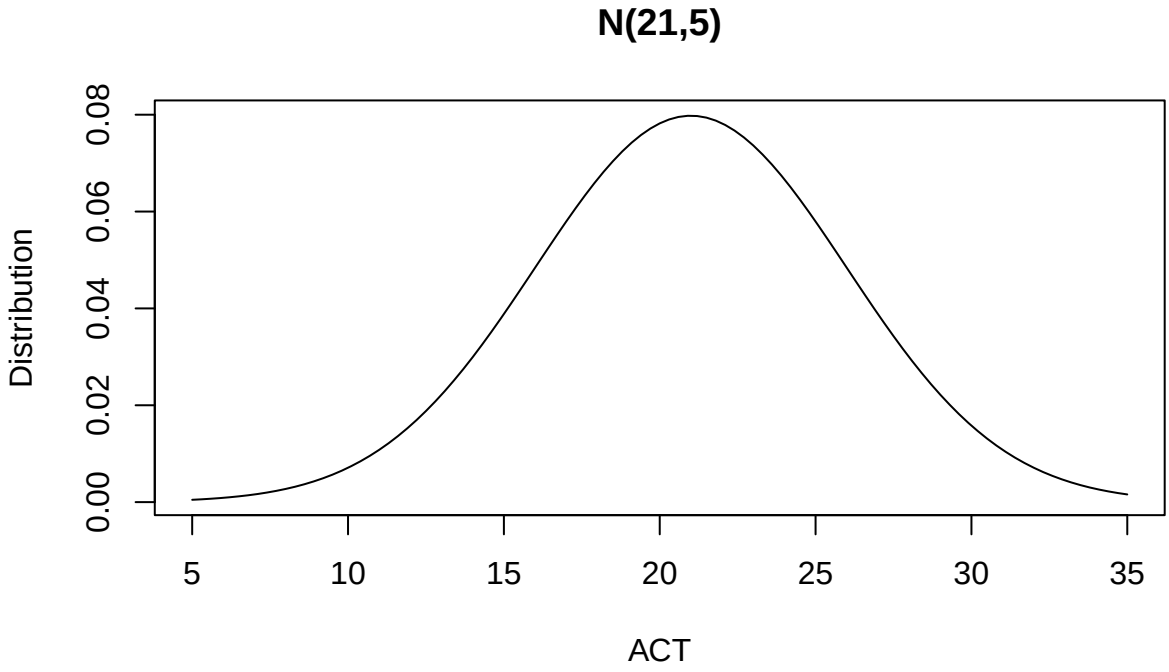
- SAT and ACT are by definition Normaly Distributed (ETS and ACT make sure this is true)

	SAT	ACT
Mean	1500.00	21.00
SD	300.00	5.00

We can plot our SAT Normal Distribution



And we can plot our ACT Normal Distribution

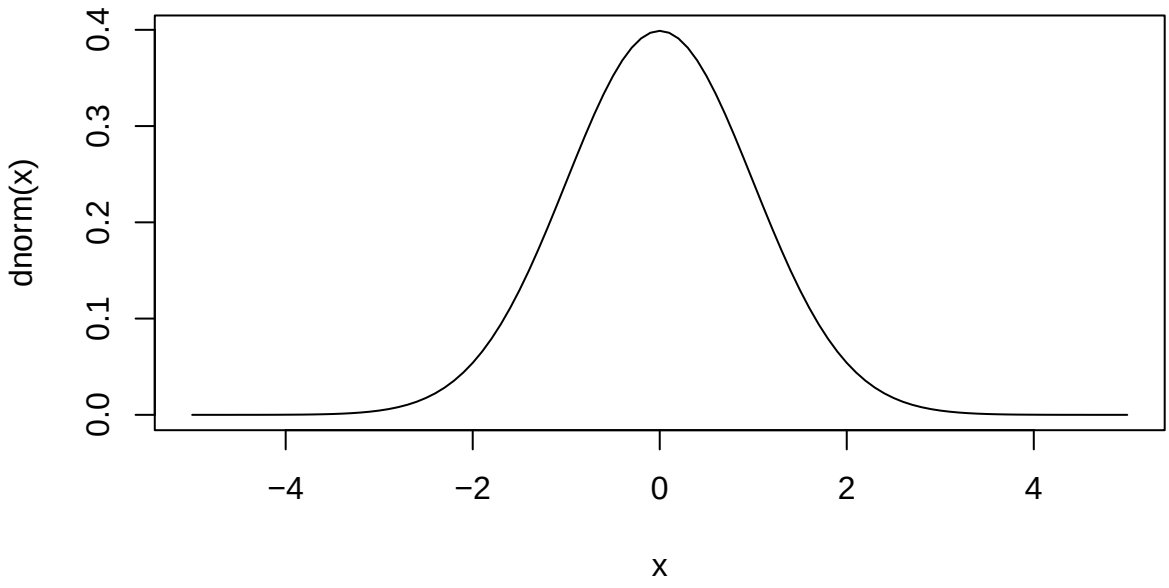


Z-Distribution (Comparing two normals)

What if we want to compare two people, Bill who has a SAT Score of 1300 and June that has an ACT score of 25? We can't do it directly! But we can transform both into a Normal Distribution that has the same mean and standard deviation (allowing us to compare them!).

Z-Distribution ($N(0,1)$)

The Z distribution is a very spetial case of the Normal distribution, it is the case where the mean equals zero and the standard deviation equals 1. I.e., $N(0,1)$



Z-score

The Z-Score is the equation for transforming normaly distributed data into the $N(0,1)$ case. Given X from a normal distribution $N(\mu, \sigma)$ we can subtract the mean to center it at 0:

$$X - \mu$$

We can then dived it by σ to normalize the standard deviation to 1 (i.e., $\sigma/\sigma = 1$):

$$\frac{X - \mu}{\sigma}$$

We call this normalized value a Z-score and represent it as, $Z = \frac{X-\mu}{\sigma}$.

Some important notes, Say we have X_1 from a $N(\mu_1, \sigma_1)$ and X_2 from a $N(\mu_2, \sigma_2)$ and we transform both into the Z-score:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \text{ and } Z_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

Then if $Z_1 > Z_2$ then we can say X_1 is of higher rank than X_2. Let's do an example with our SAT/ACT case.

ACT/SAT Example

$$Z_{Bill} = \frac{1300 - 1500}{300} = -0.67$$

and

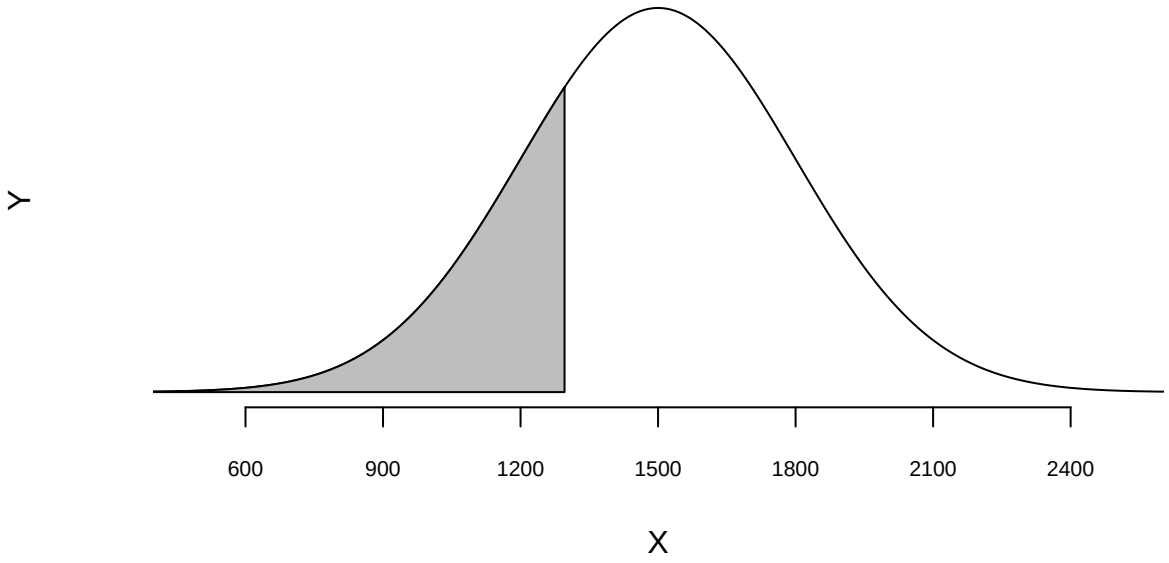
$$Z_{June} = \frac{25 - 21}{5} = 0.8$$

Thus we know know that June scored higher Bill!

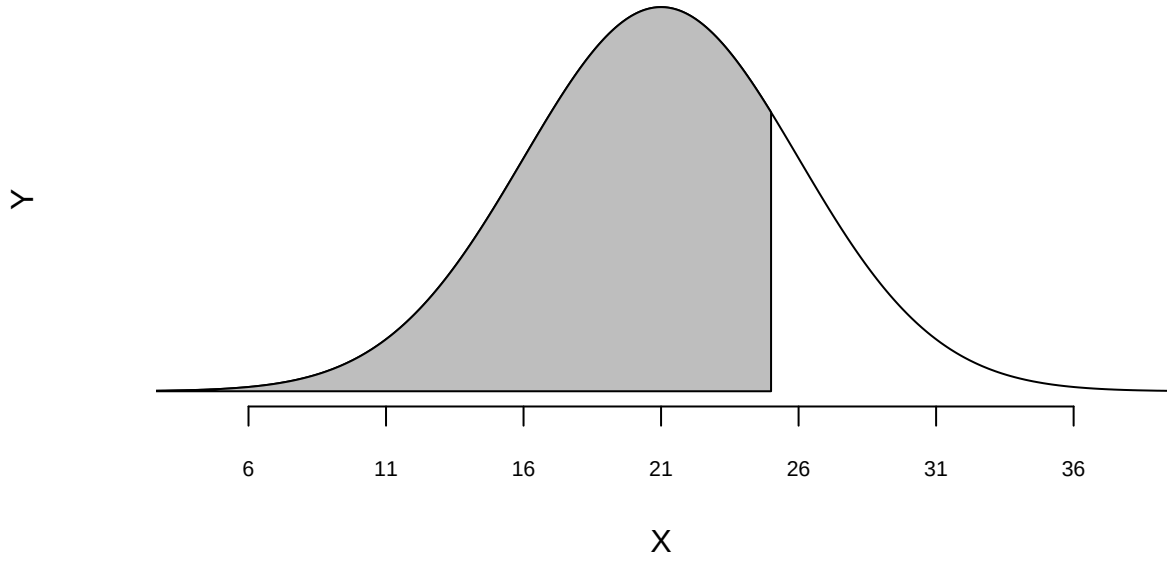
- What if we want to know their relationship to the general population of interests? I.e., what is their percentile rank?

Percentile Rank (Cummulative Distribution Function)

To find the Percentile Rank of Z-Score we need to find the area under the curve upto our Z-score (this is the Cummulative Distribution because we are summing all of the area upto the Z-score value), i.e.,



and



This area can be calulated (With R) as,

$$P(X < Z = z) = pnorm(Z_{value})$$

I.e., June has percentile rank of '$pnorm(0.8)$' = 0.788 and Bill has percentile rank of '$pnorm(-0.67)$' = 0.251.

Finding Score based on Percentile Rank

What if Bill decided he did not like how he performed and that we want to score in the top 90% of score takers?

- Step 1: Calculate the needed Z-Score, Z^*=qnorm(.9) = 1.282
- Step 2: Sove for X!

$$Z = \frac{X - \mu}{\sigma}$$

$$X = Z\sigma + \mu$$

$$X = 1.282 * 300 + 1500 = 1884.465$$

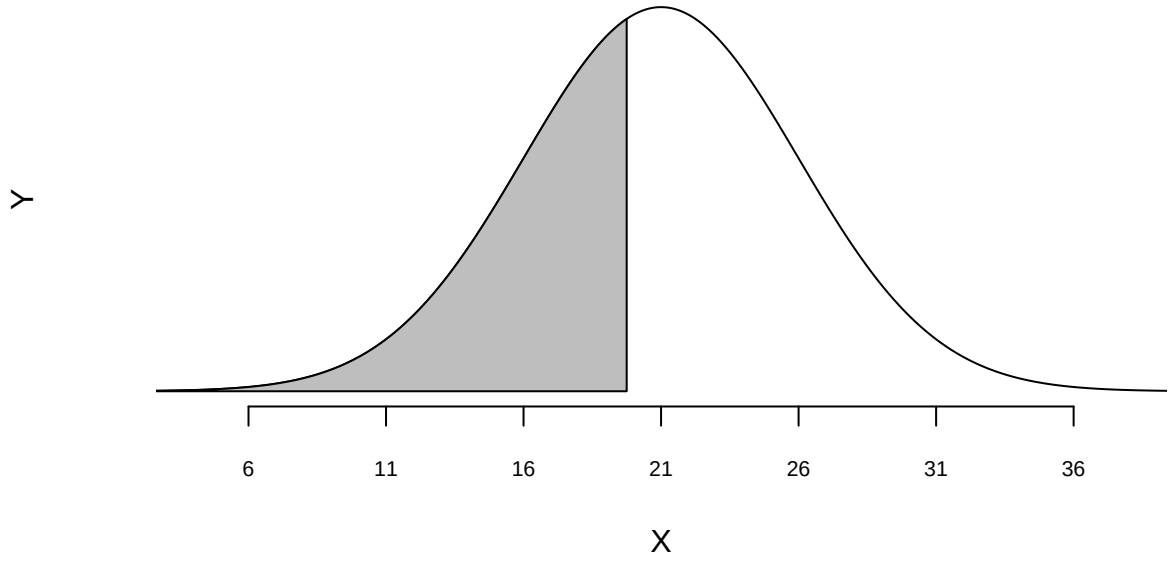
What if Bill had wanted to score in the top 2.5% of score takers? This would have been Z=1.96!

More Percentile Examples

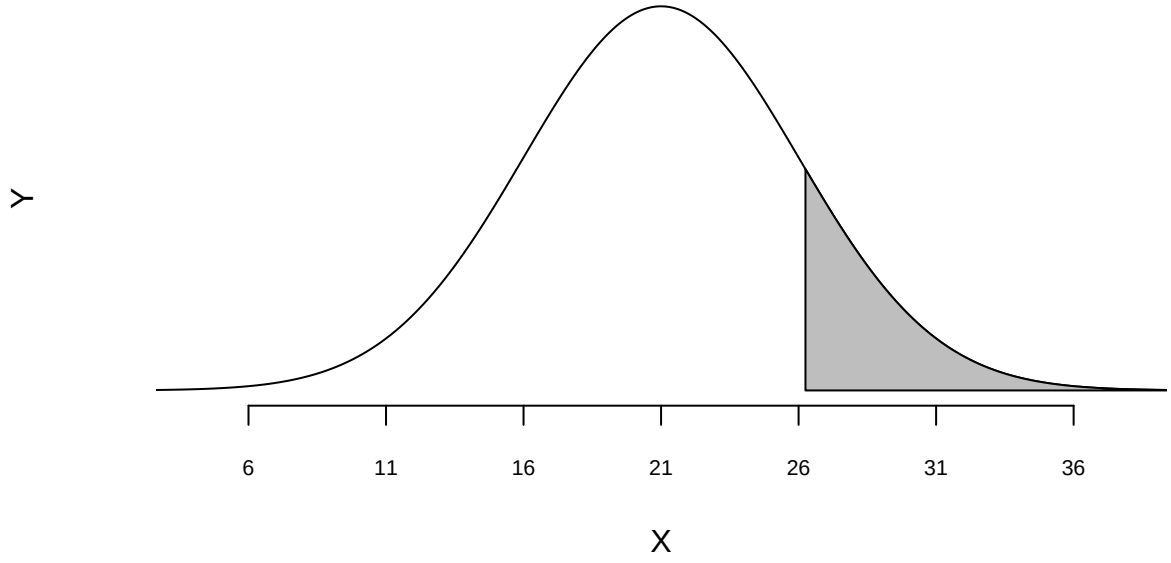
What if I run a school and I plan to accept anyone in 40% to 85% ACT score how would I figure out the range?

We can use this to find $P(Z_L^* < Z < Z_U^*)$ + First we need to find $P(Z < Z_L^*) = 0.4$ to do this we can again use `qnorm(.4)` = -0.253 to find Z_L^*,

$$X_L = Z_L^*\sigma + \mu = -0.253 * 5 + 300 = 298.733$$



+ Next I need to find Z_U^* such that $P(Z<Z_U^*)$, which is 1.036

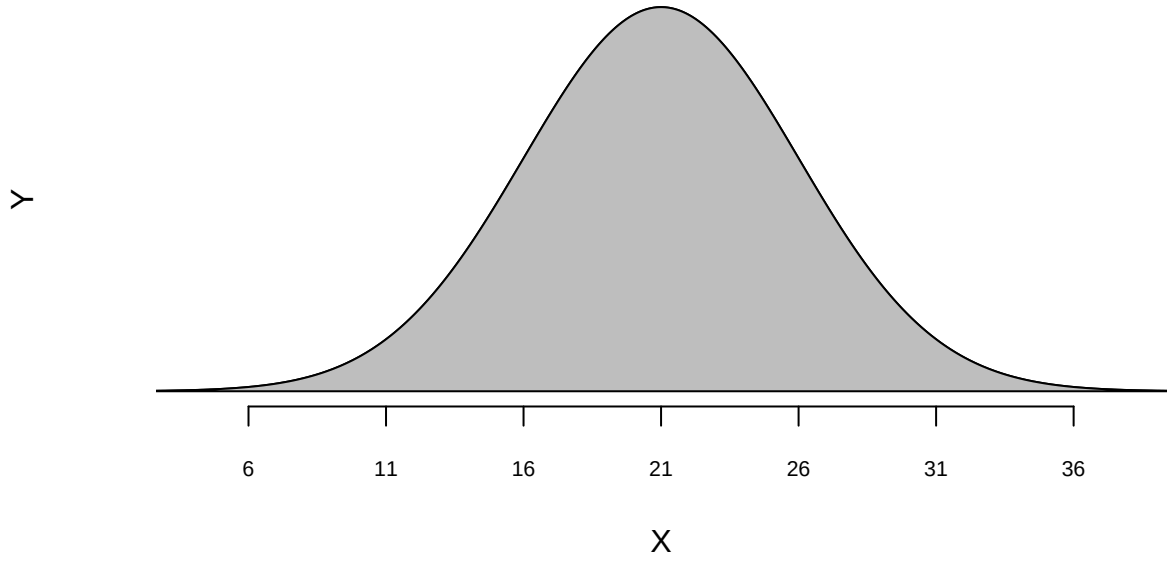


$$X_U = Z_U^* \sigma + \mu = 1.036 * 5 + 300 = 305.182$$

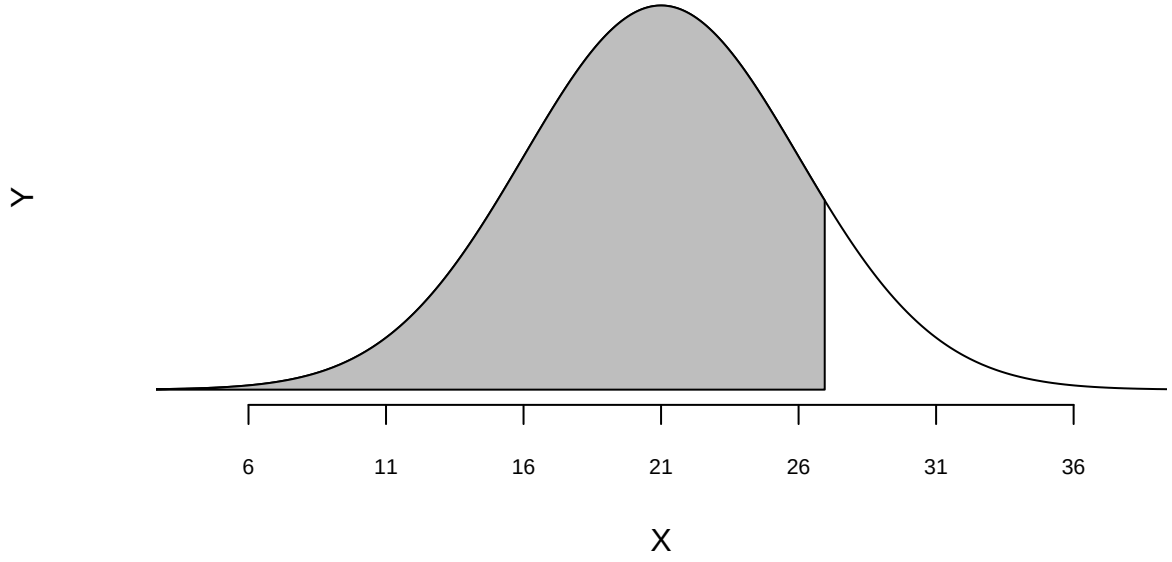
So the school will except any one from 298.733 to 305.182.

Add, Subtracting Normal Distributions

Remember, $P(Z) = 1$, i.e., the total sum of all the area under the curve is 1.



and $pnorm(Z^*)$ is the sum of all probability (area) up to Z^*, e.g. pnorm(1.2) = 0.885



Z-Test, Statistical Decision making with the Normal Distribution

Hypothesis Test with Z-Test

Define Your Hypothesis

- H_0 X is equal to VALUE.

- H_A X is less than VALUE (one sided test).
- H_A X is greater than VALUE (one sided test).
- H_A X is not equal to VALUE (two sided test).

Define Your Test

- We say H_0 is true when $X < Z_{Critival}$ and we say we reject H_0 and accept H_A when $X > Z_{Critival}$.
- We find $Z_{Critical}$ based on our chosen α-level (if not specified $\alpha = 0.05$).
  - Thus $Z_{Critical}$ is really Z_α
  - A *Critical Value* for a two sided Z-test for $\alpha = 0.05$ is 1.96

– A *Critical Value* for a one sided Z-test for $\alpha = 0.05$ is 1.65 (upper taile) or -1.65 (lower tail)

The Z-test is as follows,

Find

$$Z = \frac{X - VALUE}{\sigma}$$

Compare with Z_α.

Accept or reject Null Hypothesis.

Z-Test Example

From fivethirtyeight.com we the R package (`fivethirtyeight`) we can get the party affiliation and age of member of congress, between January 1947 and Februrary 2014. Below is the first five entries of the data set.

party	age
D	85.9
D	83.2
D	80.7
R	78.8
R	78.3

From this data we learn that age for each party is normally distributed with Democrats being N(53.431, 10.979) and Rublicans being N(53.167, 10.297).

Question Is Edward James Patten elected in 1979 at 73.4 years young for his party (Democrats)

H_0: Edward's age of 73.4 is equal to average of 53.431.

H_A: Edward's age of 73.4 is less/greater/not equal to the 53.431.

$$Z_{value} = \frac{73.4 - 53.4}{11} = 1.8$$

H_A: (Lower tail) Edward's age of 73.4 is less than 53.4

1.8 is greater than $qnorm(.05)$ so we reject H_A and accept the NUll.

H_A: (upper tail) Edward's age of 73.4 is greater than 53.4

1.8 is greater than $qnorm(.95)$ so we accept H_A and reject H_0.

H_A: (two tail) Edward's age of 73.4 is not equal 53.4.

$|Z_{value} = 1.8$ is less than $qnorm(.975)$ (two tail, so 2.5% in each tail) so we reject H_A and accept H_0. **Notice that two-tail test is most stringent – Default to two-tail test if not specified**

Statistical Significance and P-Value

P-Vaue = $P(|Z_{Value}| < Z)$, so for the last example $P(|1.8| < Z) = 2 * (1 - pnorm(abs(1.8))) = 0.072$.

We say the P-Value is *Statistically Significant* if P-Vaue $< \alpha$. So in the above example if $\alpha = 0.05$ then we would say *Not Statistically Significant*, since 0.072; however if we took $\alpha = 0.1$ then 0.072 is less than α and we say it is statistically significant.

Type 1 and Type 2 Errors

Definition Type 1 error is False Postives (i.e., you predict 1 when you should have predicted 0).

Definition Type 2 error is False Negatives (i.e., you predict 0 when you should have predicted 1).

Your α-level of significance controls your Type 1 error. I.e., your Type 1 error is lower the smaller your α level; however, the smaller your α level the larger your Type 2 errors (This is also known as the *Power of your test*). You have to balance these two issues out. We will do more with this later in the semester.

- Key take away: lower α lower type 1 error and higher type 2 error.
- When might this be acceptable: Example, medical studies. We want to know the medicine works, we want low type 1 errors!

What if our data is not from a normal distribution?

We can do this! We will do this by taking advantage of the **Central Limit Theorem (CLT)** . CLT provides us an amazing mathematical observation: The distribution of *any* mean statistic (reguardless of data form/type) given large enough sample size is normally distributed (under most reasonable conditions).

Example Mean of fatal airline accidents from 1985-1999

Let's return to our distribution of fatal airline accidents from 1985-1999:

	airline	fatal_accidents_85_99
1	Aer Lingus	0
2	Aeroflot	14
3	Aerolineas Argentinas	0
4	Aeromexico	1
5	Air Canada	0

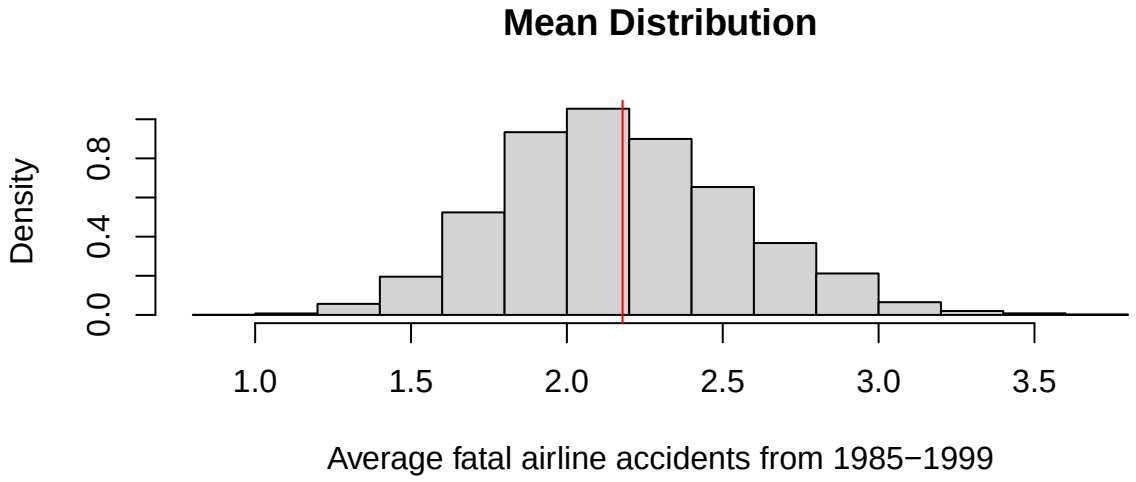
Then the distribution of fatal accidents from 1985-1999 would be the histogram of this data,

This distribution is clearly not NORMAL!!

What is the distribution of means? How can we understand that!?

- Let's think back on our simulation/permutation test.
- We can extend the idea to what is known as the *bootstrap*
  - The *bootstrap* is simple/clever idead
  - Resample the data with replacement and recalulate the statistics of interest
  - Repeat K number of times

Let's do this procedure for the mean statistic for the fatal accidents from 1985-1999 data.



So what is the key observation? - While the underlying data is clearly not Normal, the distribution of means is Normal!

How do we use CLT for Statistical Decision Making?

Z Transformation under CLT

Formally, $Z = \frac{t(x)-\mu}{SE}$. Under CLT the mean is $\bar{X}$, but $SE = \sigma/\sqrt{n}$. So our $Z_{value} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. I.e., $\bar{X}$ is distributed $N(\mu, \sigma/\sqrt{n})$.

Hypothesis test under CLT

H_0: $\bar{X}$ is equal to VALUE.

H_A: $\bar{X}$ is less than/greater than/not equal to VALUE. (one sidede/one sided/ two sided)

Example

We want to know if the average number of fatal accidents from 1985-1999 with mean 2.179 is equal to the number of fatal accidents from from 2000-2014 with 0.661 and $\sigma =$ '$sd(unlist(data[, 2:3]))$' and $n = 112$.

Hypothesis Test

H_0: The average number of fatal accidents from 1985-1999 is equal to the number of fatal accidents from from 2000-2014. ($\bar{X} = \bar{Y}$)

H_A:The average number of fatal accidents from 1985-1999 is not equal to the number of fatal accidents from from 2000-2014. ($\bar{X} - \bar{Y} \neq 0$)

Z-Value

$SE = \frac{\sigma}{\sqrt{n}} = 0.211$

$$Z_{value} = \frac{2.179 - 0.661}{0.211} = 7.182$$

Is 7.182 greater than or less then 1.96?

Conclusion We accept H_A and reject H_0.

P-value and Statistical Significance

P-value is $\Pr(Z > Z_{value}) = \Pr(Z > Z_{value}) = \Pr(Z > 7.182) = 2 * (1 - pnorm(7.182)) = 6.868 \times 10^{-13}$ which is really tiny – so we say this test is *Statistical Significant*.

95% (1-α) Confidence Intervals

Another important quality metric is the 95% Confidence Interval (or 1-α% CI). The 95% Confidence Interval is defined as follow,

$$\bar{X} \pm Z_{1-\alpha/2} \times SE = \bar{X} \pm 1.96 \times SE$$

or

$$\bar{X} - \bar{Y} \pm Z_{1-\alpha/2} \times SE$$

So in the last example the 95% CI would be

$$2.179 - 0.661 \pm 0.835 \times 0.211$$

Which is 1.341 to 1.694

- Key take aways if your $(1-\alpha)$ CI does not cover zero then you know your P-Value is less than $\alpha/2$ and therefore your test is *Statistically Significant*
- Remember, **The 95% CI guarntees that your population parameter is contained between your upper and lower values 95% of the time.** NOT THAT YOUR POPULATION PARAMETER IS BOUNDED BY THE UPPER AND LOWER VALUES (this is non-intuitive, be careful!).

Special Case of means, PROPORTIONS

Proportions are special case of mean that have very nice properties.

- The CLT kicks at fairly low n (i.e., $n > 30$)
- The SE can be derived from the bernoulli distribution and is $SE = \sqrt{p(1-p)/n}$
- Two sample SE under standard NULL is $\hat{p} = \frac{A+B}{TOTAL}$ with $SE = \sqrt{\hat{p}(1-\hat{p})/n}$

Example Congressional Representation and Age

From fivethirtyeight.com and the R package (`fivethirtyeight`) we can get the party affiliation and age of member of congress, between January 1947 and Februrary 2014. Here we have printed out the first five rows.

party	age
D	85.9
D	83.2
D	80.7
R	78.8
R	78.3

We are interested in the question "Are there more Democrats then Republicans under 40"? We have the following information:

- The number of Democrats under 40 is 1140.
- The number of Republicans under 40 is 853.
- The total number of Democrats is 10290.
- The total number of Republicans is 8274.

Hypothesis Test

H_0 The proportion of Democrats and Republicans under 40 is the same. $p_D - p_R = 0$

H_A The proportion of Democrats and Repulicans under 40 is not the same. $p_D - p_R \neq 0$

Z-value

$$Z = \frac{\hat{p}_D - \hat{p}_R - 0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

$$\hat{p}_D = \frac{1140}{10290} = 0.111$$
$$\hat{p}_R = \frac{853}{8274} = 0.103$$
$$\hat{p} = \frac{1140 + 853}{10290 + 8274} = 0.107$$

$$SE = 0.002$$

Thus,

$$Z_{value} = \frac{0.111 - 0.103}{0.002} = 3.386$$

So, $Z_{value} = 3.386$ is greater than 1.96 so we accept H_A and reject H_0.

Pvalue

The P-value is $\Pr(Z > 3.386) = 2 * (1 - pnorm(3.386)) = 7.093 \times 10^{-4}$. So again we say that the difference between the proportion of Democrats and Republicans under 40 is statistically significant at $\alpha = 0.05$ level.

95% CI

$$\hat{p}_D - \hat{p}_R \pm 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

$$0.008 \pm 1.96 \times 0.002$$

Which tells us that upper and lower bounds are 0.003 and 0.012. Which gives us a 95% chance that the true population proportion is contained within this interval.