22

recap…

Is Statistics we face an inverse of the problem we faced in Probability. While solving problems in Probability we were given all the parameters. In Statistics the parameters are unknown and must be inferred from the gathered data. We focus on the Bernoulli case.

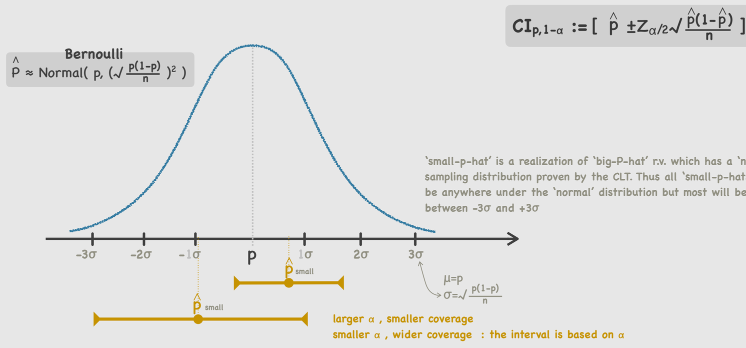
$\underbrace{X_1, \dots, X_n}_{\text{data (now is known)}} \sim \text{Bernoulli}(\, p \,)$ the unknown parameter

goals of statistical inference: you have an infinite population and you take a small random sample

1. best single guess for p is always going to be $\hat{p}$ $\Rightarrow$ $\boxed{p \approx \hat{p}}$ (Math 341 to know the other methods)
2. interval estimation (confidence interval) – range of likely values for p

$\boxed{CI_{p,1-\alpha} := [\ \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ]}$

Bernoulli $\hat{P} \approx \text{Normal}(\ p, (\sqrt{\frac{p(1-p)}{n}})^2\ )$



'small-p-hat' is a realization of 'big-P-hat' r.v. which has a 'normal' sampling distribution proven by the CLT. Thus all 'small-p-hats' can be anywhere under the 'normal' distribution but most will be between -3σ and $+3\sigma$

larger α , smaller coverage
smaller α , wider coverage : the interval is based on α

So we took 'small-p-hat' and declared the 'coverage probability' $\boxed{1-\alpha}$. i.e. if α=5% then the 'coverage probability' is $_{1-5\%=}$95% 'catches' .

All of the above is solely based on inference. Which means we do not know the 'true-p', which also means that we do not know the distribution either. Thus all we are left with is a 'small-p-hat' and its range:

Since this is all information we have, we can never verify whether we actually did 'catch' the 'true-p'. All we have is a single experiment and all we see is the 'small-p-hat'.

So what about the 'confidence interval'? Did we catch p or not? – we just can't know the unknowable!

1–α = coverage probability: how often you catch true-p inside the coverage interval
defined as probability of catching the true-p a priori to performing any actual experiments or posteriori by performing lots of experiments and examining which once did in fact catch the true-p

a smaller α produced a wider catching interval, however if the interval became too wide it produced a lot of useless information. e.g. next year Marcin will make anything between \$1 and \$1000,000. Which is true, but not precise enough to deliver any useful information. Thus making the catching interval too wide is not necessary a good strategy.

but we don't know the true-p !! 😠

what are those 'confidence intervals'

Interpretation of 'confidence intervals – CI' :

1. objective – before taking a sample the probability of 'true-p' being in CI equals 1–α

   $\boxed{P(\ p \in CI\ ) = 1-\alpha}$

   I'm going to perform only one experiment and you are telling me that I'll be catching the true-p in 95 out of a 100 experiments. Great! Your a priori information is completely useless especially that neither of us actually knows how the true-p even looks like.

2. if you take many similar samples, the number of times that 'true-p' is in CI will converge to 1–α

   $\boxed{\frac{\#\text{of } p \in CI}{n} \rightarrow 1-\alpha}$

   But this does NOT help me at all b/c in the real world I perform only one experiment (conduct one survey and I'm done) not many. Which renders this information irrelevant.

   but… P(p∈CI) after the sample is taken … = { 0,1 } I either catch p or not

   $P(\ p \in [\ \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ]\ )$ is technically an illegal statement
   no r.v.s here inside [ ]

3. if you believe in subjective probability (i.e. you are a Bayesian), under prior information you can say

   – Math 341

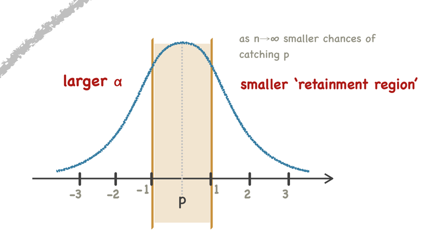
   $\boxed{P(\ p \in CI\ ) = 1-\alpha}$

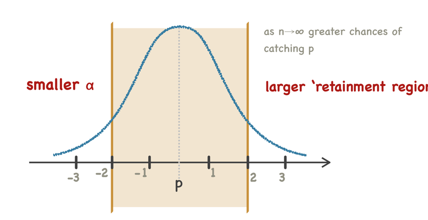
   without this assumption CIs have no meaning

   meaning the experiment is over and you firmly believe that if you were to conduct many many more such experiments the true-p will be caught 95% of the time (you assume a prior information on the true-p)

so 1–α confidence interval ≠ 1–α prob. unless you are subjective

Bayesian probability is an interpretation of the concept of probability, in which, instead of frequency or propensity of some phenomenon, assigned probabilities represent states of knowledge or belief. 📚/🙏





Hypothesis Testing testing theories about parameters

'Parameter value testing' aka 'hypothesis testing'

human-gender birth ratio:

the default position is that: P(male) = P(female) , p=50%
we call this the 'null hypothesis' denoted: $\boxed{H_0 : p = 0.5}$ (hypothesis, null)

v.s.

I believe that: P(male) ≠ P(female) , p = P(male) ≠ 50% – crazy?

the 'alternative hypothesis' is denoted: $\boxed{H_a : p \neq 0.5}$ (hypothesis, alternative)

'Occam's Razor' , 'Simple Model' – we value simple things! we do not throw out the 'null hypothesis' unless we are totally sure

$\alpha = 5\% \Rightarrow Z_{\alpha/2} = 2$

$\alpha = 1\% \Rightarrow Z_{\alpha/2} = 2.58$

R Console:

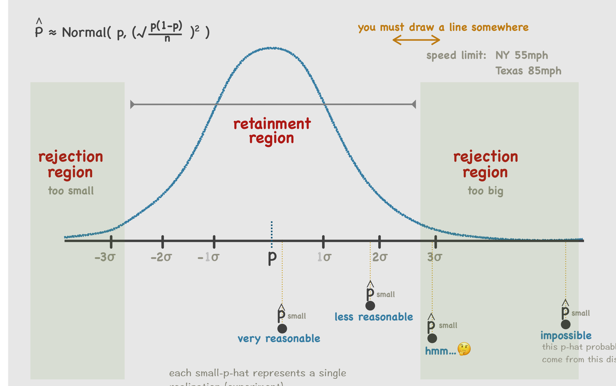
```
> qnorm(.975)
[1] 1.959964
> qnorm(.99)
[1] 2.326348
> qnorm(.995)
[1] 2.575829
```

α=5% $Z_{\alpha/2} = 2$; α=2% $Z_{\alpha/2} = 2.33$; α=1% $Z_{\alpha/2} = 2.58$

reject or retain

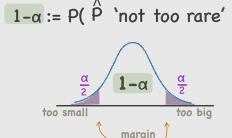
1 – take a sample of size n.
2 – assume that H_0 is true and see how 'normal' is the sample relative to H_0

$\hat{P} \approx \text{Normal}(\ p, (\sqrt{\frac{p(1-p)}{n}})^2\ )$



let $\alpha := P(\ \hat{P}$ 'too rare') very unusual

$1-\alpha := P(\ \hat{P}$ 'not too rare')



$1-\alpha = P(\ \hat{P} \in [\ \hat{p}_{small}, \hat{p}_{big}\ ]\ ) = P(\ \hat{P} \in [\ p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\ ]\ )$

$[\ p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\ ]$ – 'retainment region'

$[\ p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\ ]^C$ – 'rejection region'

now, do the experiment and calculate $\hat{p}$

if $\hat{p} \in$ 'retainment region' $\Rightarrow$ retaining H_0 , we do not have enough evidence to reject the 'null hypothesis'

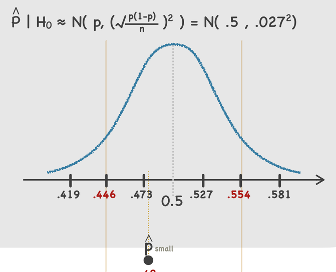
if $\hat{p} \notin$ 'retainment region' $\Rightarrow$ rejecting H_0 , we have enough evidence to reject the 'null hypothesis'

n = 345 births , males = 169
α = 5% picking α is a subjective choice
$\Rightarrow$ 1–α = 95% remember your quantiles? '68-95-997 rule' 95→[±2σ]

'retainment region' = $[\ 0.5 \pm 2\sqrt{\frac{0.5(1-0.5)}{345}}\ ] = [\ .446\ ,\ .554\ ]$ (by H_0)

$\hat{p} = \frac{169}{345} = .48$ $[\ .446 < .48 < .554\ ]$ ✔ $\hat{p} \in$ 'retainment region'

$\hat{P} \mid H_0 \approx N(\ p, (\sqrt{\frac{p(1-p)}{n}})^2\ ) = N(\ .5\ ,\ .027^2)$



retaining H_0, we do NOT have enough evidence to reject equal human-gender birth ratio.

why do we care ?

Flip a coin 100 times. Test the theory that the coin is unfair.

Fair coin means p = P(H) = 50%

scenario 1: you get 51 Heads $\Rightarrow \hat{p}$=51% fair? – Yes
scenario 2: you get 98 Heads $\Rightarrow \hat{p}$=98% fair? – No
common sense

scenario 3: you get 61 Heads $\Rightarrow \hat{p}$=61% fair? – hmm…🤔 need Math 241

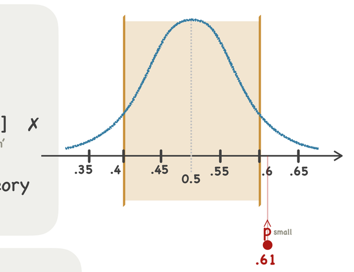
choose your α

n = 100 , heads = 61
α = 5% $\Rightarrow Z_{\alpha/2} = 2$

'retainment region' = $[\ 0.5 \pm 2\sqrt{\frac{0.5(1-0.5)}{100}}\ ] = [\ .4\ ,\ .6\ ]$

$[\ .4\ ,\ .6 < .61\ ]$ ✗ $\hat{p} \notin$ 'retainment region'

rejecting H_0, we do have enough evidence to reject the theory that the coin is fair.

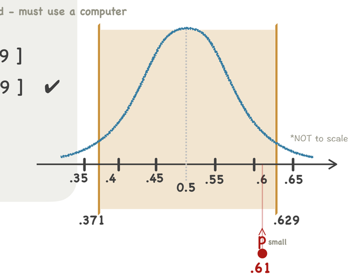


n = 100 , heads = 61
α = 1% $\Rightarrow Z_{\alpha/2} = 2.58$ remember? this can't be integrated – must use a computer

'retainment region' = $[\ 0.5 \pm 2.58\sqrt{\frac{0.5(1-0.5)}{100}}\ ] = [\ .371\ ,\ .629\ ]$

$[\ .371 < .61 < .629\ ]$ ✔ $\hat{p} \in$ 'retainment region'

retaining H_0, we do NOT have enough evidence to reject the theory that the coin is fair.



α matters

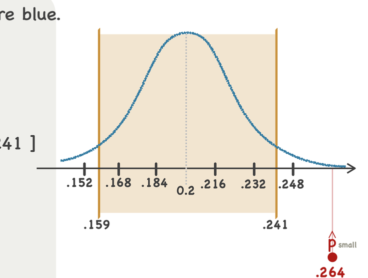
Mars (the maker of m&ms) claims that 20% of all m&ms are blue.

H_0: p = 0.2 n = 636 (total M&Ms in the bag)
H_a: p ≠ 0.2 blue = 168
α = 1% (2.58)

'retainment region' = $[\ 0.2 \pm 2.58\sqrt{\frac{0.2(1-0.2)}{636}}\ ] = [\ .159\ ,\ .241\ ]$

$\hat{p} = \frac{168}{636} = .264 \notin$ 'retainment region' $\Rightarrow$ rejecting H_0

We have enough evidence to reject the claim by the Mars candy company that the proportion of blue m&ms is 20%.



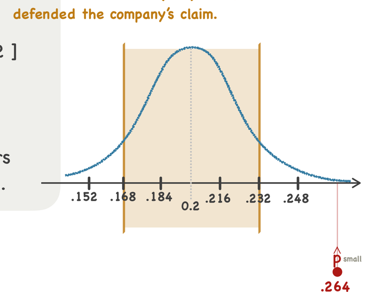
notice how the larger α gave us a stronger evidence to reject the H_0, which means that if we made a smaller (α≈0) we could have defended the company's claim.

α = 5% (2)

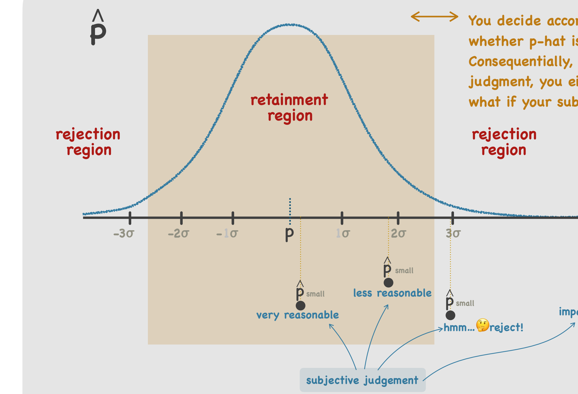
'retainment region' = $[\ 0.2 \pm 2\sqrt{\frac{0.2(1-0.2)}{636}}\ ] = [\ .168\ ,\ .232\ ]$

$\hat{p} = \frac{168}{636} = .264 \notin$ 'retainment region' $\Rightarrow$ rejecting H_0

We have enough evidence to reject the claim by the Mars candy company that the proportion of blue m&ms is 20%.



how to choose α



You decide according to your subjective analysis whether p-hat is reasonable or not. Consequentially, based on your subjective judgment, you either jettisoned H_0 or kept it. But what if your subjective judgement was wrong?

Suppose you retained a completely valid H_0. Did you make a mistake? – No
Suppose you rejected a completely valid H_0. Did you make a mistake? – Yes **type I error**
Suppose you retained a completely wrong H_0. Did you make a mistake? – Yes **type II error**
Suppose you rejected a completely wrong H_0. Did you make a mistake? – No

	retained	rejected
H_0 true	✔	✗ type-I error
H_0 false	✗ type-II error	✔

P(type I error) = P(reject | H_0 true) = α (you chose α)

P(type II error) = P(retained | H_0 false) = ⋯ (not in this course)

P(rejected | H_0 false) = POWER

allows you to find new theories. if α is too small you never find anything new. Don't be overly skeptical.

if you make α too big, you risk the possibility of type I error
α↑ $\Rightarrow$ P(type I error)↑ $\Rightarrow$ P(type II error)↓ but the possibility of type II error goes down
α↓ $\Rightarrow$ P(type I error)↓ $\Rightarrow$ P(type II error)↑
if you make α too small, you risk the possibility of type II error; but the possibility of type II error goes up

Court Case in America

H_0 – innocent until proven guilty
H_a – guilty until proven innocent
Decision: punish or not

type I error: you punish an innocent person → Cost ?
if you choose big α , everyone goes to jail

type II error: you let a guilty person walk → Cost ?
if you choose small α , criminals walking the streets

the Costs will dictate what α should be

Alarm System

H_0 – NO Fire
H_a – Fire
Decision: alarm

type I error: false alarm → Cost ?
if you choose big α , you have an oversensitive alarm
- hurt ears but no fires
- no-one takes the alarm seriously

type II error: Fire but no alarm → Cost ?
if you choose small α , your alarm never goes off
- your house burned down

Clinical Trial

H_0 – drug does NOT work
H_a – drug works
Decision: release drug

type I error: release a drug which does NOT work to the market → Cost ?
if you choose big α , there will be lots of ineffective drugs overflowing the market
- lawsuits
- deaths

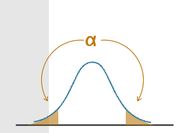
type II error: failed to release a working drug to the market → Cost ?
if you choose small α , lots of effective drugs never see the light of day
- wasted research money
- deaths

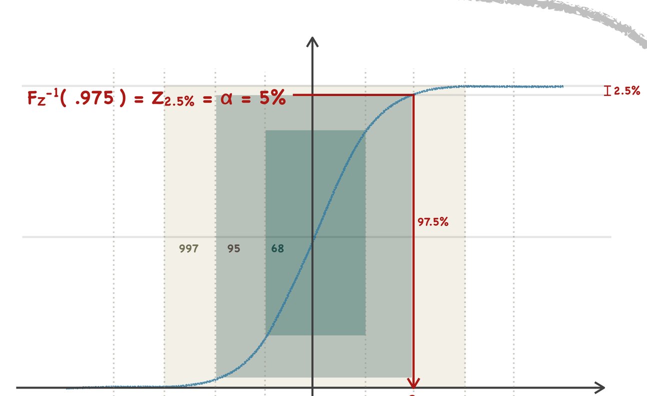
R Console and data visualizations

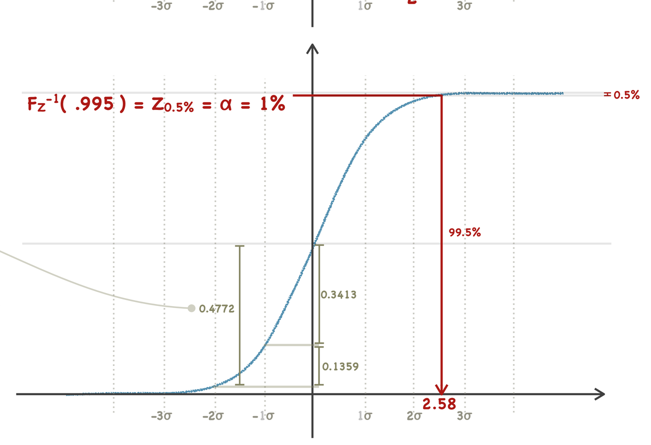
- area 1 to +∞
  `pnorm(1, lower.tail = F)`
  `0.1586553`
- area -∞ to -1
  `pnorm(-1, lower.tail = T)`
  `0.1586553`
- get difference between adjacent quantiles e.g. area between -2 and -1
  `diff(pnorm(-2:-1))`
  `0.1359051`
- area under normal curve from -2 to 2
  `sum(diff(pnorm(-2:2)))`
  `0.9544997`
- area -∞ to -3
  `pnorm(-3, lower.tail = T)`
  `0.001349898`
- area -3 to -2
  `diff(pnorm(-3:-2))`
  `0.02140023`

```
> qnorm(.975)
[1] 1.959964
> qnorm(.99)
[1] 2.326348
> qnorm(.995)
[1] 2.575829
```

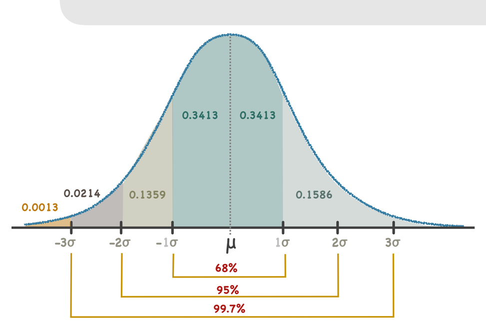
α=5% $Z_{\alpha/2} = 2$; α=2% $Z_{\alpha/2} = 2.33$; α=1% $Z_{\alpha/2} = 2.58$

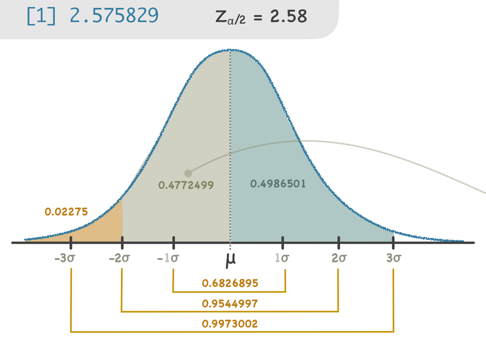


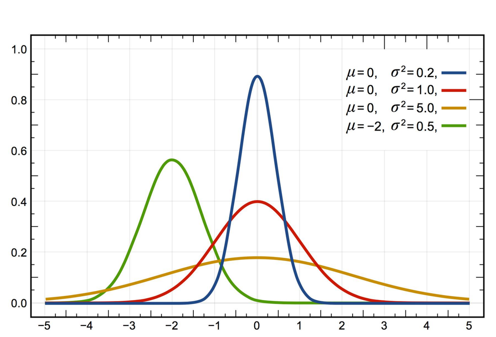


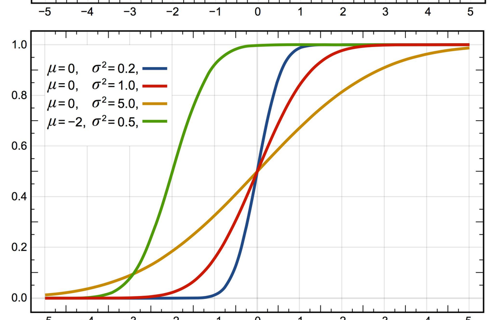


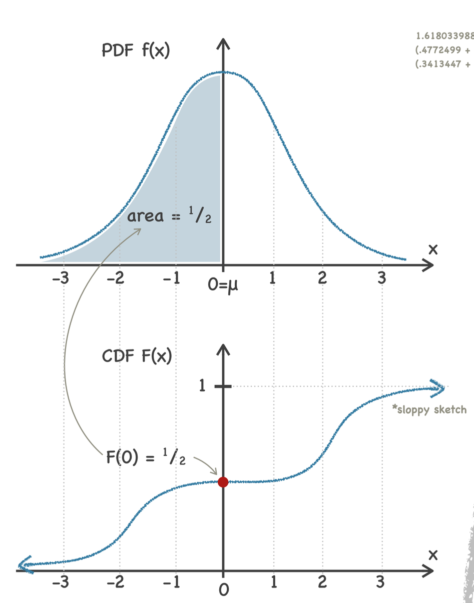
1.618033988… = φ
(.4772499 + .3413447) / .4772499 = 1.715233
(.3413447 + .1359051) / .3413447 = 1.398146

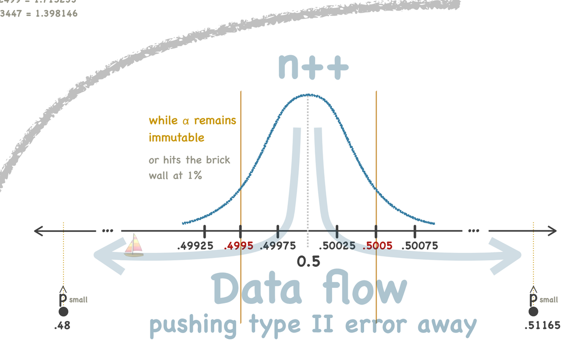












In order to be more confident that our interval actually contains the population mean (true-p) we can increase the confidence interval of our estimates—from 95 to 99, for example. This forces us to become less precise. How can we mitigate that tradeoff between level of confidence and the precision of our interval? We do this primarily by increasing our sample size – n.

Sample Size (n) — Larger samples result in smaller standard errors, and therefore, in sampling distributions that are more clustered around the population mean (true-p). A more closely clustered sampling distribution indicates that our confidence intervals will be narrower and more precise.

As the sample size increases, the interval and its width decrease, thus providing a more precise estimate of the population value.

To reduce the width of a confidence interval by a factor of two (i.e., in half), you have to quadruple the sample size.

23

tinkering with n – data size

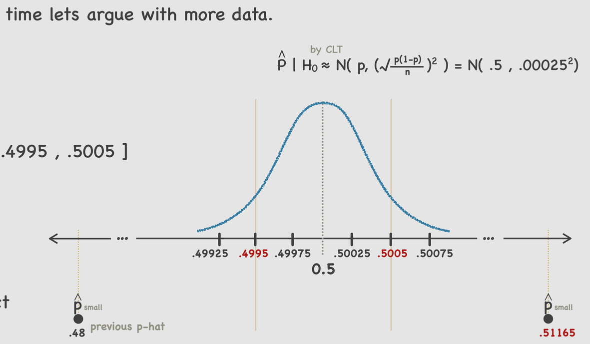
lets do another birth ratio experiment, but this time lets argue with more data.

n = 4,247,000 births , males = 2,173,000
degree of skepticism α = 5%

$\hat{P} \mid H_0 \approx N(\ p, (\sqrt{\frac{p(1-p)}{n}})^2\ ) = N(\ .5\ ,\ .00025^2)$ (by CLT)

'retainment region' = $[\ 0.5 \pm 2\sqrt{\frac{0.5(1-0.5)}{4,247,000}}\ ] = [\ .4995\ ,\ .5005\ ]$ (by H_0)

$\hat{p} = \frac{2,173,000}{4,247,000} = .51165 \notin$ 'retainment region'



rejecting H_0 , we have enough evidence to reject that male and female are born equally likely

Increasing the size of our data while simultaneously keeping the same 'retainment region' has greatly added to the precision of our experiment. We can clearly see that .51165 is way of the chart which substantiates the rejection of equal male to female birth ratio hypothesis.
Furthermore, if we were to decrease the precision of our experiment by decreasing α to 1% thus concomitantly increasing the 'retainment region', our empirical .51165 still remains way off.

thus our previous experiment with n=345, allowed for 'type II error' to creep in.
However by increasing the data size to n=4,247,000 we successfully managed to push 'type II error' out and debunk the 'null hypothesis'.

$\Downarrow$

n↑ $\Rightarrow$ α no change $\Rightarrow$ P(type II error)↓ $\Rightarrow$ POWER↑ (α: P(type I error))

n↑ $\Rightarrow$ P(type II error)↓
large n can mitigate the precision diffusive effects of reducing α

All science works this way. We have a theory and we keep it until we see enough evidence to prove it otherwise. We don't throw good theories away willy-nilly though. We tend to remain conservative by keeping α relatively small.

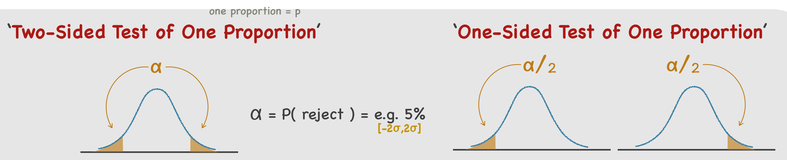
another look at α

H_0 : UFOs and aliens do NOT exist 👽
H_a : UFOs and aliens do exist

if α-small , H_0 supporter: skeptical, impossible to persuade, pigheaded, no matter what type of evidence you present, you just can NOT persuade him/her that aliens do exist.

if α-large , H_0 supporter: overly open-minded, susceptible to persuasion, gullible, you can come up with a shady evidence and the supporter changes his mind and start believing in aliens.

'Two-Sided Test of One Proportion' (one proportion = p) **'One-Sided Test of One Proportion'**



α = P(reject) = e.g. 5% [-2σ,2σ]

recall $Z_{\alpha/2} := F_Z^{-1}(\ 1-\frac{\alpha}{2}\ )$ $F_Z^{-1}(\ .975\ ) = Z_{2.5\%} = \alpha = 5\%$

α = P(reject) = P(rejection region)
1–α = P(retain) = P(retainment region)
= P(Z ≤ zα)

$P(\ \sqrt{\frac{p(1-p)}{n}}\,Z \le z_\alpha\sqrt{\frac{p(1-p)}{n}}\ )$ · both sides by σ_Z

$P(\ p + \sqrt{\frac{p(1-p)}{n}}\,Z \le p + z_\alpha\sqrt{\frac{p(1-p)}{n}}\ )$ + p to both sides

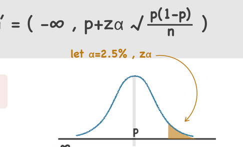
$P(\ \hat{P} \le p + z_\alpha\sqrt{\frac{p(1-p)}{n}}\ ) \Rightarrow$ 'retainment region' = $(\ -\infty\ ,\ p + z_\alpha\sqrt{\frac{p(1-p)}{n}}\ )$

de-standardization: multiply by S.E. add the mean $\Leftarrow$ ($\mu = p$, $\sigma = \sqrt{\frac{p(1-p)}{n}}$) standardization: subtract the mean and divide by S.E. (recall)

standardization: $\hat{P} \xrightarrow{\text{by CLT}} \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \longrightarrow \frac{\hat{P} - p + p}{\sqrt{\frac{p(1-p)}{n}}} \cdot \sqrt{\frac{p(1-p)}{n}} = \hat{P}$

de-standardization: $Z \longrightarrow Z \cdot \sqrt{\frac{p(1-p)}{n}} + p = \hat{P}$

recall $Z_{\alpha/2} := F_Z^{-1}(\ 1-\frac{\alpha}{2}\ ) \Rightarrow \boxed{Z_\alpha := F_Z^{-1}(\ 1-\alpha\ )}$



Über example

'Right-Sided Test of One Proportion'

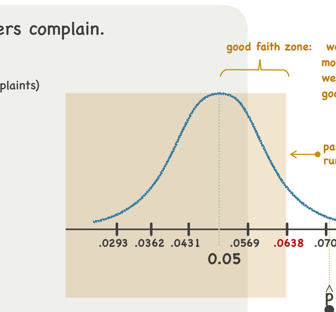
Über fires drivers if more than 5% of passengers complain.

H_0 : good driver if p ≤ 5% (less then or equal to 5% complaints) — inequality → one-sided test (RHS)
H_a : bad driver if p > 5% (more than 5% complaints)

Decision: fire or keep after 1000 evaluations

p = 5% = P(complaint)

experiment unveils 71 complains



'retainment region' = $[\ (-\infty\ ,\ 0.05) + 2\sqrt{\frac{0.05(1-0.05)}{1000}}\ ] = [\ -\infty\ ,\ .0638\ ]$ (by H_0)

$\hat{p} = \frac{71}{1000} = .071 \notin$ 'retainment region'

rejecting H_0 , we have a sufficient evidence to reject the 'null hypothesis' that the driver is good

Imagine the driver has 60 complaints which makes the little p-hat = 6%

Since 0.05 ≠ 0.06 , Über should fire the driver. The driver clearly has more than 6% of complaints but according to our experiment still remains employed. This seems like a contradiction.
Why 0.05 = 0.06 ? b/c Über only cares about the real inherent probability of a complaint the 'true-p' which is unknowable. Remember we can only see the little p-hat, and we construct the region around it hoping to catch the 'true-p'. Thus Über wants to be absolutely sure that the 'true-p' is within that catching range before it takes any drastic actions such as firing a driver. Über chooses to remain conservative about rejecting H_0 and belies in the abilities of their drivers while in the same time giving them the benefit of a doubt (good faith zone).

Über example 2

'Left-Sided Test of One Proportion'

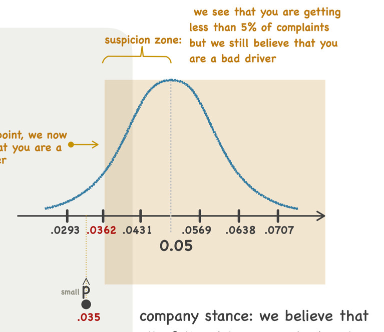
Über fires drivers if more than 5% of passengers complain.

H_0 : bad driver if p ≥ 5% (greater then or equal to 5% complaints) — inequality → one-sided test (LHS)
H_a : good driver if p < 5% (less than 5% complaints)

Decision: fire or keep after 1000 evaluations

p = 5% = P(complaint)

experiment unveils 35 complains



'retainment region' = $[\ (0.05\ ,\ +\infty) - 2\sqrt{\frac{0.05(1-0.05)}{1000}}\ ] = [\ .0362\ ,\ +\infty\ ]$ (by H_0)

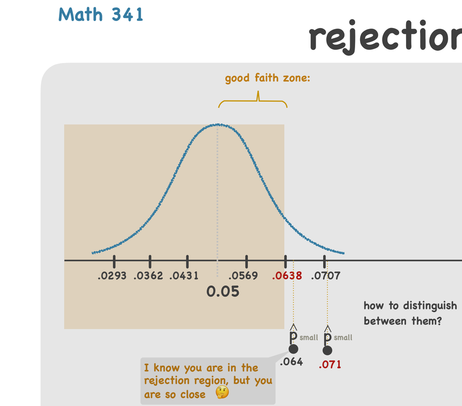
$\hat{p} = \frac{35}{1000} = .035 \notin$ 'retainment region'

rejecting H_0 , we have a sufficient evidence to reject the 'null hypothesis' that you are a bad driver

company stance: we believe that all of the drivers are bad, unless they really convince us otherwise.

Math 341

rejection strength



$P_{val} := P(\ \hat{p}$ or more extreme | H_0 true)

$= P(\ \hat{P} \ge \hat{p} \mid H_0\ )$

$= P(\ \hat{P} \ge \hat{p} \mid p=0.05\ )$

$= P(\ \hat{P} \ge 0.071 \mid p=0.05\ )$

$= P(\ \hat{P} \ge 0.071 \mid p=0.05\ ) = P(\ \frac{\hat{P} - .05}{.00689} \ge \frac{.071 - .05}{.00689}\ ) \approx P(\ Z \ge 3.05\ )$ (probability statement; standardization: subtract the mean and divide by S.E.; by CLT)

≈ 0.001144207

$\approx .1\ \% < \alpha = 2.5\%$

very clearly below 2.5%
– this represents the strength of rejection