- Probability

Eg. spam detection (twitter bot)

R.V. E.g. sample a tweet at random, is it spam or not?

$X = \{0, 1\}$ 0 : if not spam
1 o.w.

Prob. Distributions

P: Values X can take $\rightarrow [0, 1]$

(a) $P(X = x) \geq 0$ for all values x R.V. X can take

(b) $\sum_{\text{values } X \text{ can take}} P(X = x) = 1$

What is $P(X=1)$?

(a) the prob. that randomly sampled tweet is spam $\Rightarrow$

(b) The proportion of tweets that are spam in the set of "all" tweets

The oracle of TWEET:

Suppose we are omniscient and know that 5% of "all" tweets are spam.

What is $P(X=1) \stackrel{?}{=}$

What is $P(X=0) \stackrel{?}{=}$

→ what if I randomly chose n=100 tweets, How many of those do I expect to be spam?

Expectation

$$EX = \sum_{\substack{\text{values} \\ X \text{ can take}}} x \cdot p(X=x)$$

What is the expectation of X?

$1(.05) + 0(.95) = .05$

What is the expectation of $Y = X_1 + X_2 + \cdots + X_n$

Well assuming X_i's are independent

$$EY = E[X_1 + X_2 + \cdots + X_n] = EX_1 + EX_2 + \cdots + EX_n$$

$$= .05 + .05 + \cdots + .05 = 100(.05) = 5$$

Note also: $E\,aX = a\,EX$ (a constant)

$Y = 100X \Rightarrow EY = 100\,EX = 100(.05) = 5$

All of this assumed we know $p(X=1) = .05$, but we don't. We need to estimate it.

So, we have data $x_1, x_2, x_3, \ldots, x_{100}$ (suppose 7 of those tweets are labeled as spam).

so $y = \sum x_i = 7$, now we expect $y = np$

so let's use that to estimate p

$$np = 7 \Rightarrow 100p = 7 \Rightarrow \hat{p} = \frac{7}{100} = .07$$

It's wrong but close, but is it any good?

A couple of observations:

① Our estimate of p ($\hat{p}$) is the sample mean of $X_1, X_2, \ldots, X_{100}$

Let's do an experiment and sample many times (plots in R)

Obs.

① The distribution of $\hat{p}$ is centered at p

② The spread of the distribution depends on n the number of samples

Two central tenets of statistics

① law of large numbers (LLN)

Given data $X = X_1, \ldots, X_n$ with $EX = \mu$

(a) $EY = \frac{1}{n}\sum_i X_i = \mu$

(b) $\mathrm{Var}\, Y \to 0$ as $n \to \infty$

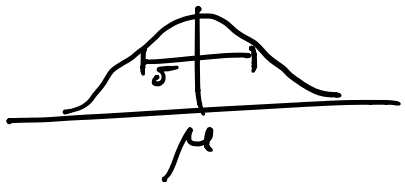
"under some assumptions"

② Central Limit theorem

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \text{a } \underline{\text{normal}} \text{ distribution}$$

as $n \rightarrow \infty$

What is a normal distribution?



ⓐ symmetric around mean parameter μ

ⓑ exponential decay with rate σ

$$\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Remember that bit about probability summing to 1? This is equivalent to saying area under curve = 1. To make satisfy that we need

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

This is the normal density function with mean μ and variance σ^2

We write that as $N(\mu, \sigma)$.

Notice the term inside the square?

$z_i = \left(\frac{x_i - \mu}{\sigma}\right)$ this is the standardization transformation we saw before! So, taking a R.V. that is normalby distributed with mean μ, standard deviation σ (i.e. $X \sim N(\mu, \sigma)$) and we apply the same transform $\frac{X-\mu}{\sigma} = Z$ to get new R.V. Z, we say that $Z \sim N(0,1)$. We call $N(0,1)$ the standard normal distribution.

② CLT (continued)

Now we can finish this statement

$Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, ⓐ $EY = EX$ (LLN), ⓑ s.d. $Y = \frac{s_x}{\sqrt{n}}$, ⓒ $Y \sim N\left(EX, \frac{s_x}{\sqrt{n}}\right)$ as $n \to \infty$.

Equivalently, $$\frac{Y - EX}{\frac{S_X}{\sqrt{n}}} \to N(0,1)$$

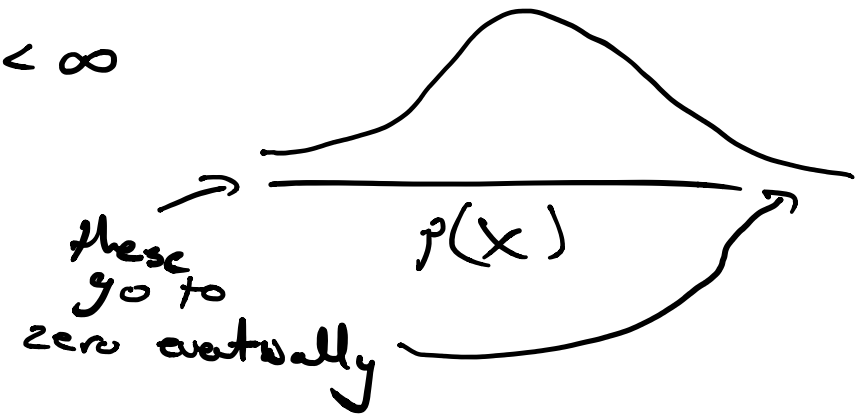
as $n \to \infty$

Disclaimer: There are a lot of mathematical subtleties here. Two important ones

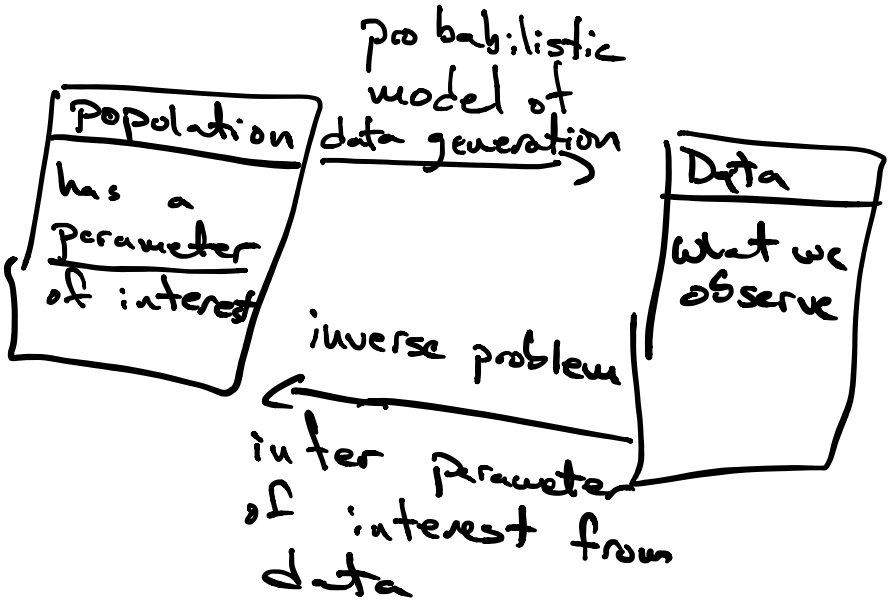
(a) $X_1, \ldots, X_n$ are iid

i: independent
id: identically distributed

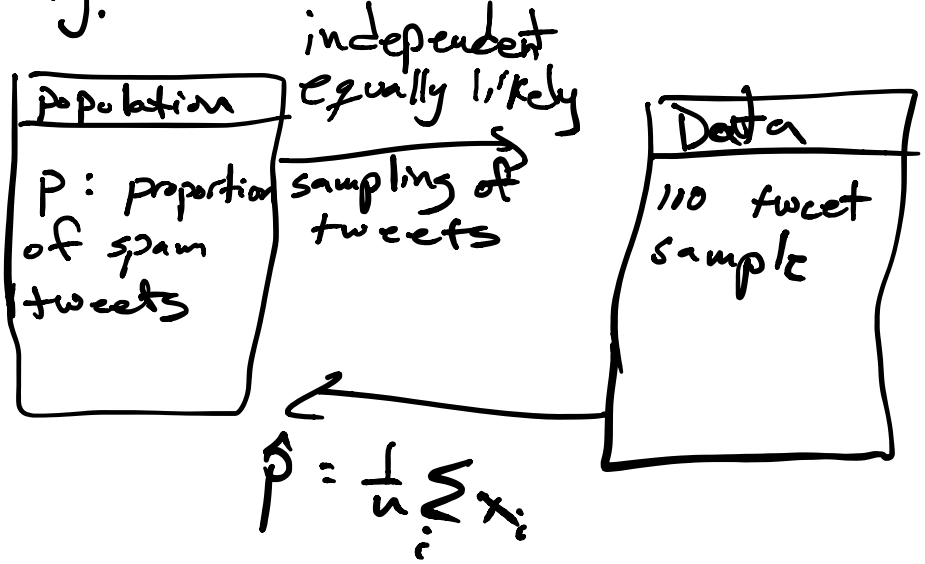
(b) $\mathrm{Var} X < \infty$



OK, so our schematic description of inference is



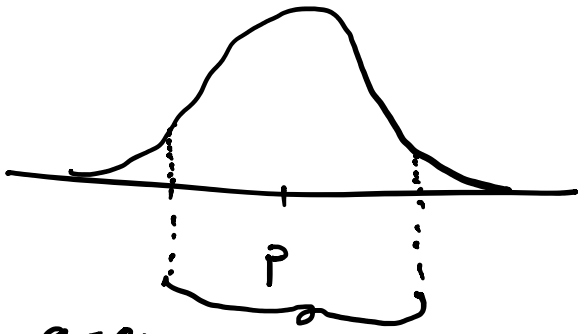
E.g.



Now we can answer, how confident are we?

well,

$$\hat{p} = \frac{1}{n}\sum x_i \sim N\left(p, \frac{\sigma}{\sqrt{n}}\right)$$



with 95% probability we should see $\hat{p}$ in this range! (we call this confidence interval)

But, what is σ??

$$\sqrt{\text{var}(X)}$$

$$\text{var}(X) = E\left[X - EX\right]^2$$

$$= \sum_{\substack{\text{all values} \\ X \text{ can take}}} (x - EX)^2 \times P_r(X = x)$$

$$= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p$$

$$= p^2(1-p) + (1-p)^2 \cdot p$$

$$= p^2 - p^3 + (1 - 2p + p^2)p = p^2 - p^3 + p - 2p^2 + p^3$$

$$= p - p^2$$
$$= p(1-p)$$

So, $\text{Var } X = p(1-p)$

We have an estimate!!

$$\frac{\hat{\sigma}}{\sqrt{n}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

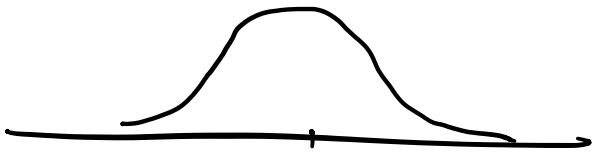
so, our estimate of the proportion of spam tweets given sample $x_1, \ldots, x_n$ is

$$\hat{p} \pm qnorm\left(.95, \hat{p}, \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)$$

Hypothesis Testing

How else is this framework useful? Suppose that before I sampled tweets I thought (hypothesized) that 50% of tweets are spam.

Under this hypothesis, estimates $\hat{p}$ would be distributed as



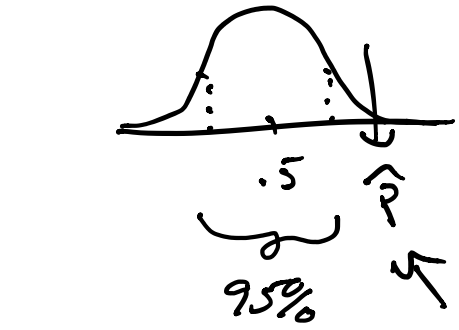
$$N\left(.5, \frac{\sqrt{.5(1-.5)}}{\sqrt{n}}\right)$$

Once we have our sample, we can <u>test</u> this hypothesis:

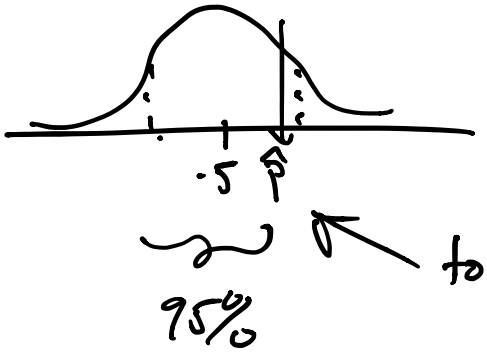
H_0: $p = .5$ (null)

H_1: $p \neq .5$ (alternative)

If we see $\hat{p}$ (sample mean from our sample of tweets is <u>too far</u> from .5 then we could <u>reject</u> our hypothesis. How can we say it's too far? Probability!! if $Pr(|.5 - |\hat{p}|) > .95$, we say it's too far and we reject.



↖ too far, reject!



↖ too close can't reject.

Now, this 95% rejection threshold is conservative but somewhat arbitrary. So we use one more metric $Pr(|v| = \hat{p})$ (the infamous p-value) to say: we would reject this hypothesis for all thresholds greater than this p-value.

Summary:

Inference: estimate parameter from data based on e.g. expectation from assumed probability model.

For <u>averages</u> the LLN and the CLT tells us how to compute probabilities from our <u>single</u> parameter estimate

With these we can provide confidence intervals for our estimate

Testing: Having a hypothesis about our parameter of interest, we can use probability under <u>this hypothesis</u> to see how data agrees with hypothesis and reject it if it doesn't agree enough

Distributions:

In this example we saw three distributions:

Bernoulli(p): $X \in \{0, 1\}$

parameter: $p = Pr(X = 1)$

$EX = p$

$\text{var}X = p(1-p)$

Binomial(n, p): $Y \in \{0, 1, 2, \ldots\}$

$Y = \sum_{i=1}^{n} X_i$ where X_i are iid Bernoulli(p)

parameters: $p = Pr(X_i = 1)$

n

$EY = np$

$\text{var}Y = np(1-p)$

Normal (Gaussian) $N(\mu, \sigma)$:

$X \in \mathbb{R}$

parameters: μ: center of distribution

σ: decay rate

$EX = \mu$

$\text{Var}X = \sigma^2$

Useful references:

https://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/
https://www.openintro.org/stat/textbook.php

For a majority of these distributions, R has the

- d___ : density function
- p___ : probability function
- q___ : quantile function
- r___ : random value generator

family of functions

For Binomial(n, p):

- dbinom(x, n, p)
- pbinom(x, n, p)
- qbinom(p, n, p)
- rbinom(num values, n, p)

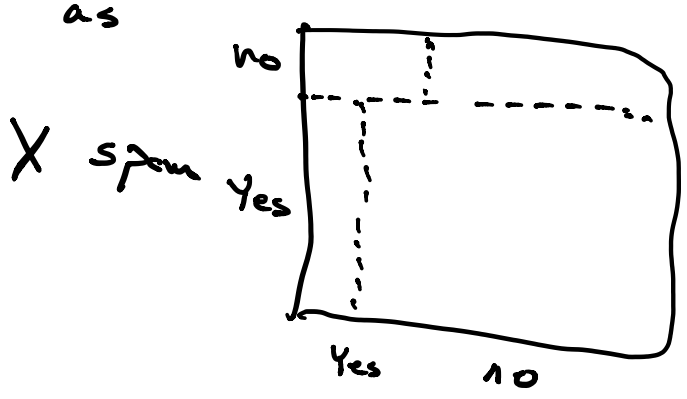
Joint and conditional probability

Suppose that for each tweet I observe I can say if it has a lot of retweets or not. So, I have another variable:

$$Y = \begin{cases} 1 & \text{lots of retweets} \\ 0 & \text{otherwise} \end{cases}$$

(Note, we could say $Y \sim \text{Bernoulli}(p_Y)$)

So, we could think of the population of all tweets as



We can talk of the joint probability distribution of X and Y

$\Pr(X = x, Y = y)$

Here we have the same conditions:

$$Pr(X=x, Y=y) > 0 \quad \text{for all combinations } x \text{ and } y$$

and

$$\sum_{\substack{\text{all combinations} \\ x, y}} Pr(X=x, Y=y) = 1$$

We can also talk about conditional probability where we look at the probability of a tweet being spam conditioned of it not having lots of retweets:

$$Pr(X=x \mid Y=y)$$, now to make this a probability distribution we need to make sure

$$\sum_{\substack{\text{all values} \\ X \text{ can take}}} Pr(X=x \mid Y=y) = 1$$

We do that by writing $Pr(X=x \mid Y=y) = \frac{Pr(X=x, Y=y)}{Pr(Y=y)}$

With this we can also talk about independence: if the probability of spam does not depend on a tweet having lots of retweets then

$$Pr(X=x) = Pr(X=x/Y)$$

we say they are independent. Refer to the tweet population diagram, are X and Y independent there? What would the diagram look like if X and Y were independent?

With conditional probability, we can get conditional expectation

$$E[X/Y=y] = \sum_{\substack{\text{all values} \\ X \text{ can take}}} x \; Pr(X=x/Y=y)$$

With conditional probability and expectation in hand, we can do Machine Learning!!