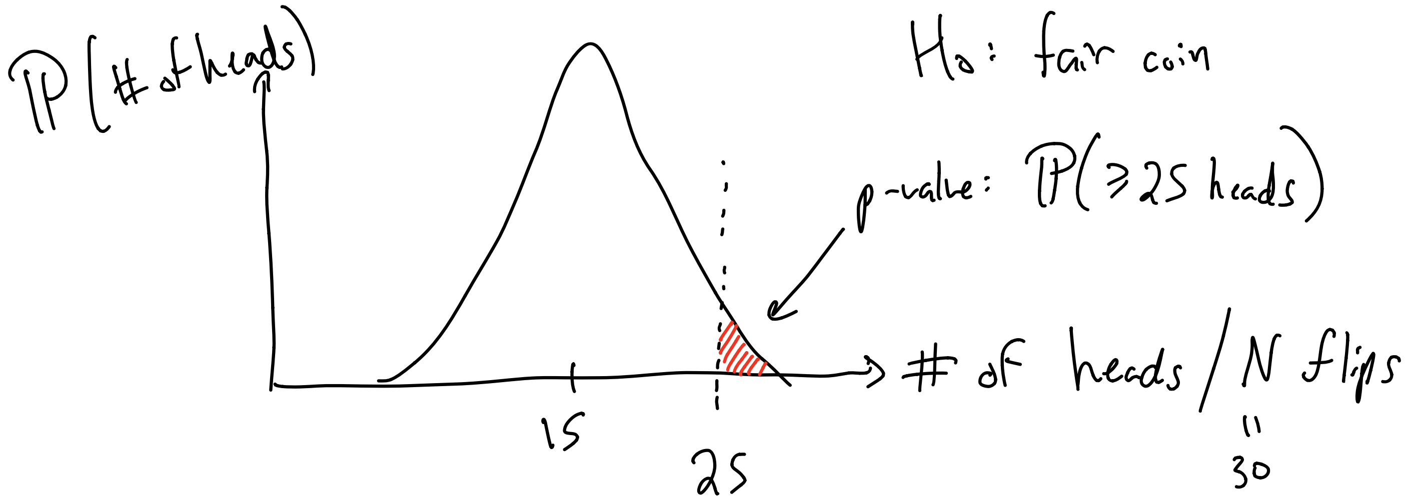
Week 6 Section

Danylo L Week 6
MCB 112

Some of what we discussed in lecture:

→ <u>given</u> some null hypothesis on how data is generated, how surprising is a particular observation of data?



For section today, a mix of things but they're all connected...

→ given data, how do we make a <u>new</u> hypothesis?

→ revisit Bayes Rule

→ estimation of parameters for binomial & normal

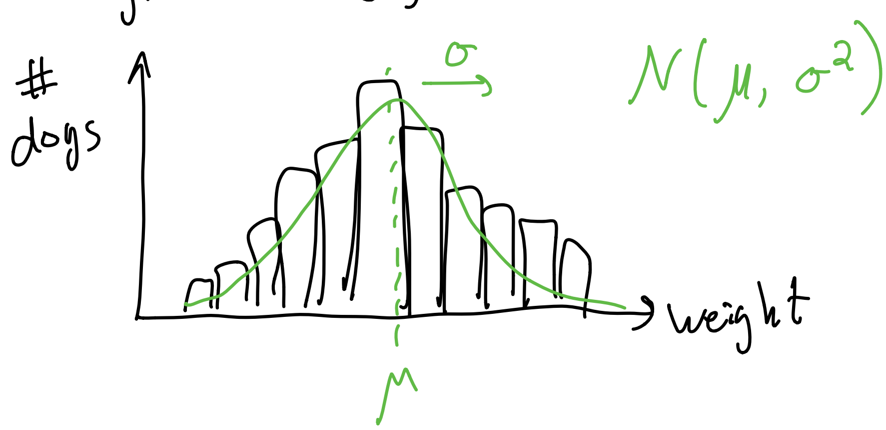
→ see how T scores arise from estimation of ↑

→ concepts behind p-set

Parameter Estimation / Inference

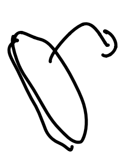
Given some data, $X_1, \dots, X_N$, we'd like to describe the process that produced the data.

Ex. 1: Weights of Siberian huskies



$\Rightarrow$ what particular μ, σ describe $X_1, \dots, X_n$?

Ex. 2: # of heads out of N flips of a coin?



$\left.\begin{array}{l} \text{H w/ prob } p \\ \text{T w/ prob } 1-p \end{array}\right\} \rightarrow N \text{ times}$

$\Rightarrow$ what particular p describes # of heads / flips?

Notation: θ is a <u>hypothesis</u> on how the data was generated

↳ particular values of the parameters of the underlying probability distribution.

<u>Ex.</u> for dog weights: $\theta = \{\mu = 15 \text{ kg}, \ \sigma = 5 \text{ kg}\}$

<u>Ex.</u> for an unbiased coin: $\theta = \{p = 0.5\}$

The big question we'd like to answer:

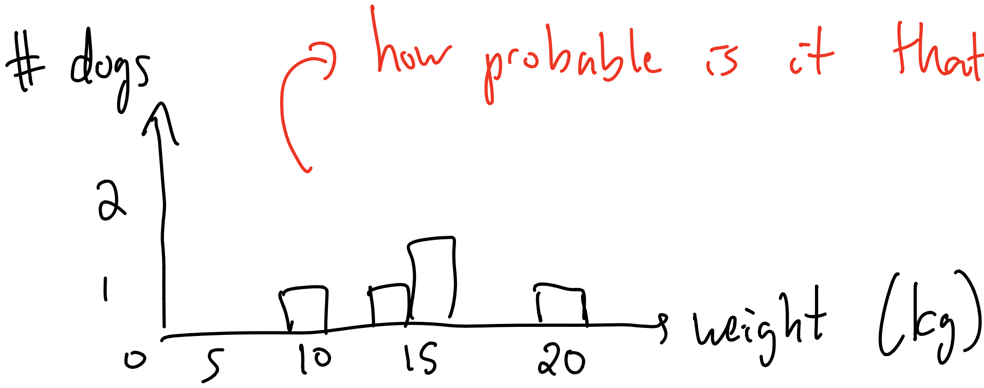
<u>given data D, what's the most probable θ?</u>

(θ: a hypothesis on how data D was generated)

<u>Ex.</u> I weigh 5 dogs and make a histogram:

how probable is it that $\mu = 5$? meh.

$\mu = 15$? probable!



<u>Ex.</u> I get 25 heads in 30 flips

↳ "fair" hypothesis: how probable is it that $p = \mathbb{P}(\text{heads}) = 0.5$?

Suppose we had some set of hypotheses, $\theta_1, \ldots, \theta_M$

Given data, how probable is a particular hypothesis θ_k?

likelihood:
how probable is the data given our hypothesis is true

prior:
how probable is our hypothesis prior to seeing data?

$$\mathbb{P}(\theta_k \mid D) = \frac{\mathbb{P}(D \mid \theta_k)\, \mathbb{P}(\theta_k)}{\mathbb{P}(D)}$$

posterior:
how probable is our hypothesis θ_k post seeing the data?

marginal:
how probable is the data under all possible hypotheses?

Marginal recap:

$$\mathbb{P}(D) = \sum_{i=1}^{M} \mathbb{P}(D \text{ AND } \theta_i \text{ true})$$

add over all possible hypotheses

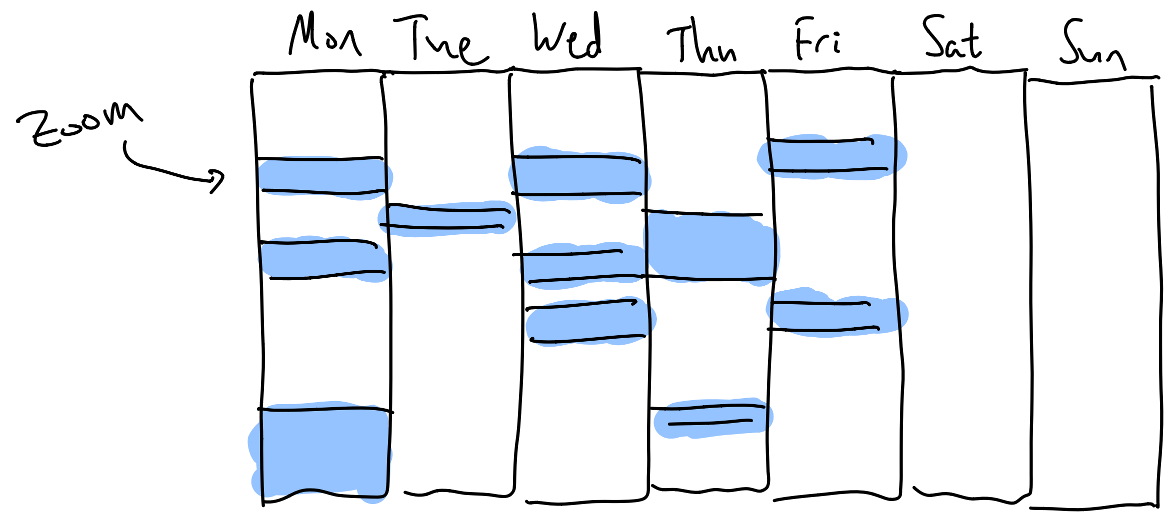
$$= \sum_{i=1}^{M} \mathbb{P}(D \mid \theta_i)\, \mathbb{P}(\theta_i)$$

$\mathbb{P}(X \text{ AND } Y) = \mathbb{P}(X \mid Y)\, \mathbb{P}(Y)$

$$= \int \mathbb{P}(D \mid \theta)\, \mathbb{P}(\theta)\, d\theta$$

if an ∞ number of θ's

An example: My Weekly Schedule



θ : What day it is. Monday, Thursday, who knows....

Data: how many Zoom meetings I had today.

$\theta \xrightarrow{\text{generates}}$ data, but also data $\xrightarrow{\text{infer}} \theta$

Ex. Marginal:

$$
\begin{aligned}
\mathbb{P}(3 \text{ meetings}) &= \mathbb{P}(3 \text{ meetings AND it's Monday}) \\
&\quad + \cdots + \\
&\quad \mathbb{P}(3 \text{ meetings AND it's Sunday}) \\
&= \sum_{i=1}^{7} \mathbb{P}(3 \text{ meetings AND it's } i^{th} \text{ day of week}) \\
&= \sum_{i=1}^{7} \mathbb{P}(3 \text{ meetings} \mid i^{th} \text{ day})\, \mathbb{P}(i^{th} \text{ day})
\end{aligned}
$$

Comparing two posteriors / hypotheses

$$\mathbb{P}(\text{Monday} \mid 3 \text{ meetings}) = \frac{\mathbb{P}(3 \text{ meetings} \mid \text{Mon})\,\mathbb{P}(\text{Mon})}{\mathbb{P}(3 \text{ meetings})}$$

$$\mathbb{P}(\text{Thursday} \mid 3 \text{ meetings}) = \frac{\mathbb{P}(3 \text{ meetings} \mid \text{Thu})\,\mathbb{P}(\text{Thu})}{\mathbb{P}(3 \text{ meetings})}$$

To compare these two, I can take a ratio.

The denominator cancels out!

Back to our q:

given data D, what's the most probable θ?

We can scan over a lot of θ's to look for a particular θ w/ the highest $\mathbb{P}(\theta \mid D)$,

$$\mathbb{P}(\theta \mid D) = \frac{\mathbb{P}(D \mid \theta)\,\mathbb{P}(\theta)}{\cancel{\mathbb{P}(D)}}$$

(can ignore denom when comparing posteriors generally)

If we have some prior beliefs $\mathbb{P}(\theta)$,

(it just feels like a Thursday...)

the θ that maximizes $\mathbb{P}(D|\theta)\,\mathbb{P}(\theta)$ is the maximum a posteriori (MAP) estimate.

If we further assume uniform priors on θ...

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\,\cancel{\mathbb{P}(\theta)}}{\cancel{\mathbb{P}(D)}}$$

then... the most probable θ given the data is the θ that maximizes the likelihood

$$\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)$$

This is **Maximum Likelihood Estimation!**

Concept behind likelihood $\mathbb{P}(D|\theta)$ w/ binomial

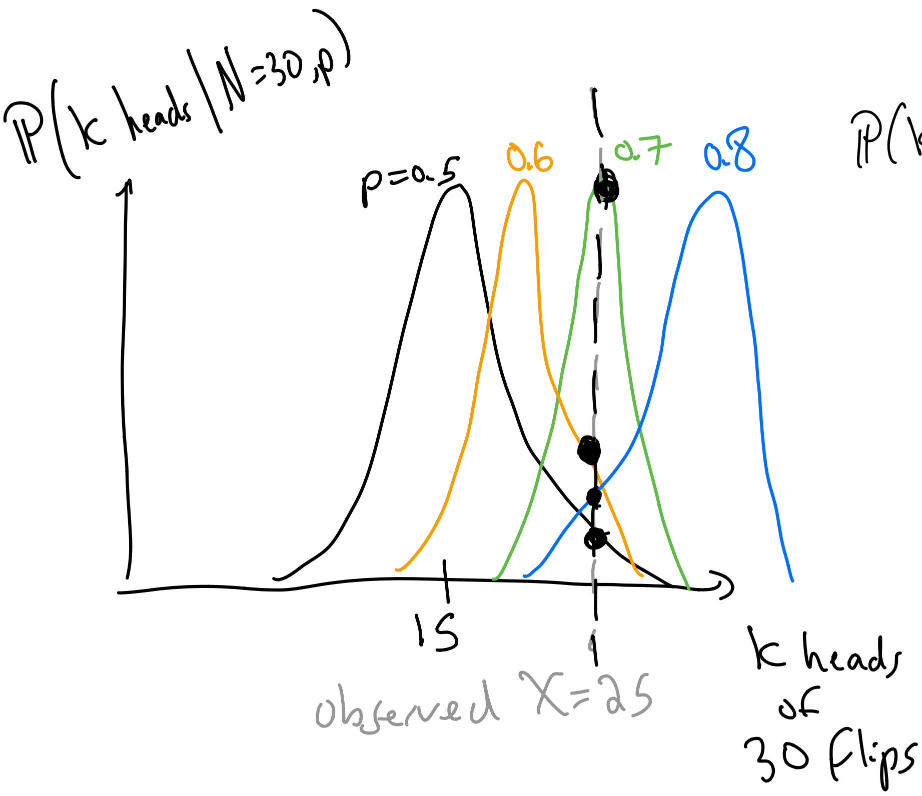
binomial process: N trials

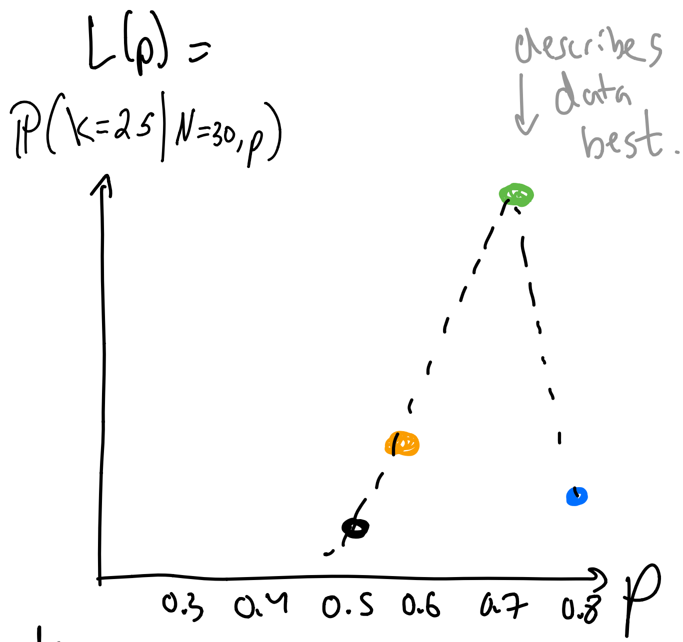
each trial has success prob. p

how many successes out of N trials?

$$\underbrace{\overbrace{\underline{p}\ \ \underline{p}}^{k \text{ success}}\ \ \underline{1-p}\ \ \underline{1-p}\ \ \underline{1-p}\ \ \underline{1-p}}_{N \text{ trials}}$$

$$\mathbb{P}(X = k \text{ successes} \mid N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$





Key point: the likelihood measures overlap of data w/ a data generation/probability process parameterized by θ.

Max likelihood of Binomial

Observed data: 25 heads / 30 flips.

What's the max. likelihood estimate of p?

The p that maximizes the likelihood.

$$\mathbb{P}(k \mid N, p) = \mathcal{L}(p) = \binom{N}{k} p^k (1-p)^{N-k}$$

This is a function of p. We can maximize it by finding p s.t. $\frac{\partial}{\partial p} \mathcal{L}(p) = 0$.

$$\frac{\partial}{\partial p} \mathcal{L}(p) = \frac{\partial}{\partial p} \left[\binom{N}{k} p^k (1-p)^{N-k} \right]$$

gross... chain rule... take the log!

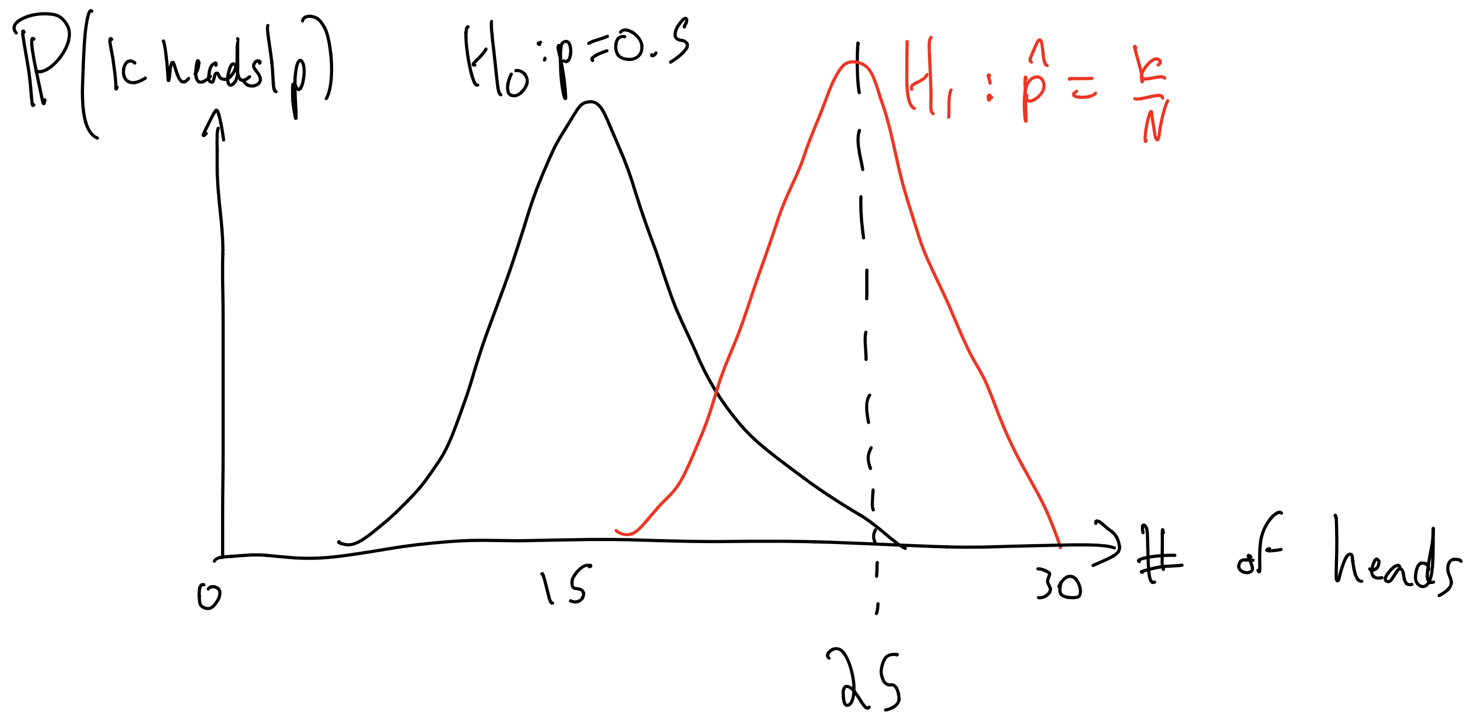
$$\log \mathcal{L}(p) = \log \binom{N}{k} + k \log p + (N-k) \log (1-p)$$

$$\frac{\partial}{\partial p} \log \mathcal{L}(p) = \frac{k}{p} - \frac{N-k}{1-p} = 0$$

$$\Rightarrow \frac{k}{p} = \frac{N-k}{1-p}$$

$$\Rightarrow k - kp = Np - kp \Longrightarrow \hat{p} = \frac{k}{N}$$

So, given we observe k heads out of N flips, the max. likelihood estimate of p is $\frac{k}{N}$.



Things to consider:

- is H_0 invalidated?
- does this mean H_1 has to be the correct model?
- what if we had prior beliefs on

$$\mathbb{P}(\theta) = \mathbb{P}(\text{prob. of heads})?$$

(maybe we really trust the coin is fair)

MLE for a normal: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

For one data point,

$$\mathbb{P}(X = x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For n data points, write the joint likelihood:

$$\begin{aligned}
L(\mu, \sigma^2) &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid \mu, \sigma^2) \\
&= \mathbb{P}(X_1 = x_1 \mid \mu, \sigma^2) \cdot \dots \cdot \mathbb{P}(X_n = x_n \mid \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \mu, \sigma^2) \\
&= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}$$

Take the log:

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}$$

↳ we'll find MLE again by taking derivatives

MLE for μ:

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2) = \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2} = 0$$

$$\Rightarrow \boxed{\hat{\mu} = \sum_{i=1}^{n} x_i \Big/ n}$$

The MLE for μ is just the sample mean.

MLE for σ^2:

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

skipping some steps... $\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$

But wait... if we only observe $X_1, \ldots, X_n$, we don't know μ!

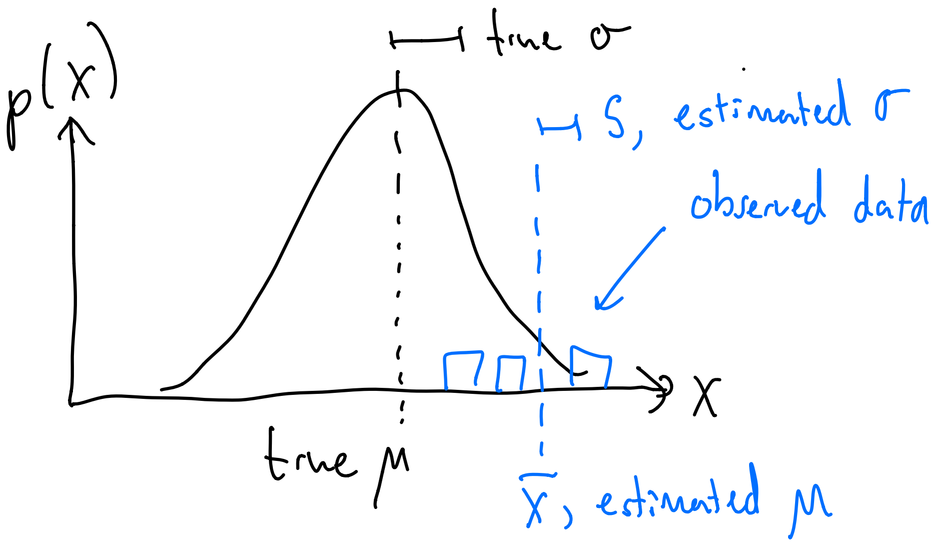
Replacing μ w/ $\bar{X}$: $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

It turns out this is a biased estimator of the population σ^2. (see Bessel's correction)

Unbiased estimate of population variance: $\boxed{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$

Hypothesis testing on $\bar{X}$

$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma)$, n small



Q: How likely is it that μ is the population mean?

Compute distance b/w μ and $\bar{X}$, scaled by typical fluctuation in $\bar{X}$, which is $\frac{\sigma}{\sqrt{n}}$. ("standard error" of the mean)

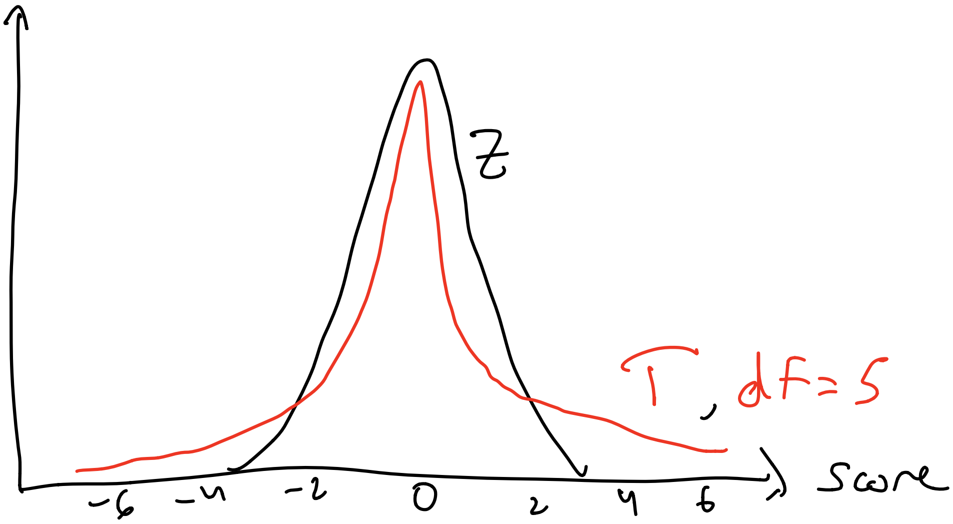
$\Rightarrow$ Standard score: $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

If σ is known, this is a Z score, $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

If σ is unknown, we estimate it with S, this is a T score: $T = \dfrac{\bar{X} - \mu}{S / \sqrt{n}}$, w/ -1 degrees of freedom

Distribution of T, Z scores

(when μ really is population mean)



$\Rightarrow$ T distribution has fatter tails.
It allows for $\bar{X}$ to be "far away" from μ because we had to estimate σ^2 from the data.

$\Rightarrow$ We could have estimated too small a σ^2, which means the T score would be larger than it should be $\rightarrow$ hence more probability of big T scores

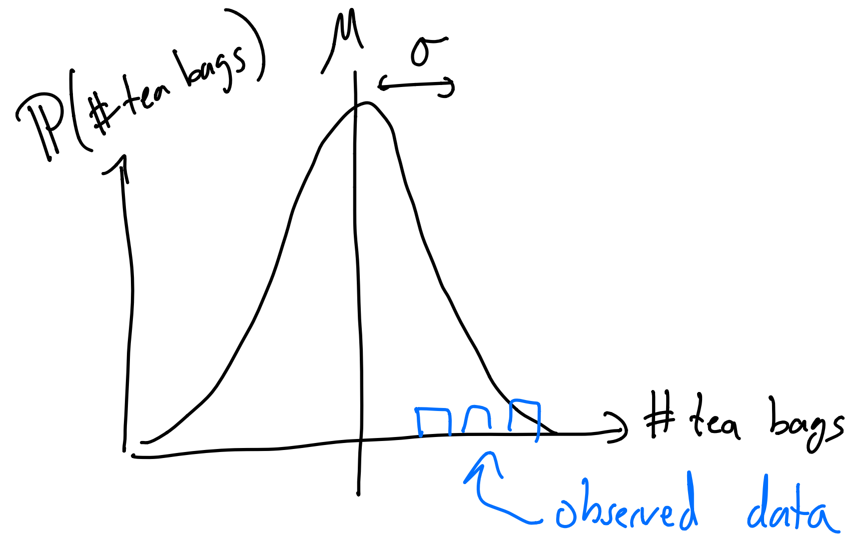
$\Rightarrow$ At large n, we estimate σ^2 well, T dist converges to Z dist.

So the T dist can arise from estimation of an uncertain σ^2 given data.

On the p-set we will marginalize over many potential σ^2's.

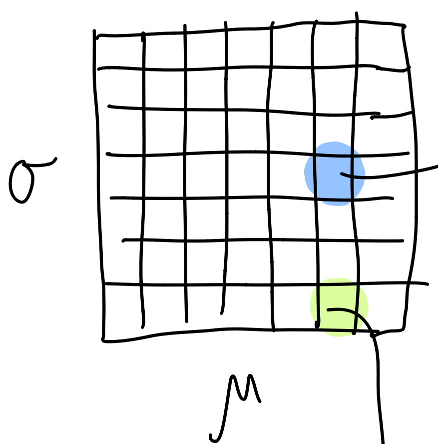
Concept for P-Set

There's a true $\mathcal{N}(\mu, \sigma^2)$, but we only see a few data points



$\Longrightarrow$ Goal: <u>bet</u> on candidate (μ, σ^2) given only a few observed datapoints

How? We'll assume a <u>grid</u> of possible (μ, σ):



one of these (μ, σ) pairs generated the data we observe, $X_1, \dots, X_n$

for any candidate (μ, σ^2) we can compute how probable they are given our data...

with a posterior, $\mathbb{P}(\mu, \sigma^2 \mid X_1, \dots, X_n)$

$$\mathbb{P}(\mu, \sigma^2 \mid X_1, \ldots, X_n)$$

$$= \frac{\overset{①}{\mathbb{P}(X_1, \ldots, X_n \mid \mu, \sigma^2)} \; \overset{②}{\mathbb{P}(\mu, \sigma^2)}}{③ \; \mathbb{P}(X_1, \ldots, X_n)}$$

① likelihood: (independent observations)

$$\mathbb{P}(X_1, \ldots, X_n \mid \mu, \sigma^2) = \mathbb{P}(X_1 \mid \mu, \sigma^2) \times \ldots \times \mathbb{P}(X_n \mid \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \mathbb{P}(X_i \mid \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}$$

② prior: $\mathbb{P}(\mu, \sigma^2) = \mathbb{P}(\mu)\,\mathbb{P}(\sigma^2)$

③ marginal: $\mathbb{P}(X_1, \ldots, X_n) = \sum_{\mu} \sum_{\sigma} \mathbb{P}(X_1, \ldots, X_n \text{ AND } \mu, \sigma)$

$$= \sum_{\mu} \sum_{\sigma} \mathbb{P}(X_1, \ldots, X_n \mid \mu, \sigma)\,\mathbb{P}(\mu, \sigma)$$

Extra: distribution of a sample mean.

The expectation, or average, of $\bar{X}$:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i]$$

(mean of sum = sum of means)

$$= \frac{1}{N}(N\mu)$$

$$= \mu.$$

So the average $\bar{X}$ is indeed μ, the population average.

Now for the spread in $\bar{X}$? Its variance:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)$$

$$= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right)$$

$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X_i)$$

independent variables

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2$$

Each $X_i \sim N(\mu, \sigma^2)$

$$= \frac{1}{N^2}N\sigma^2$$

$$= \frac{\sigma^2}{N}$$

$$\mathrm{SD}(\bar{x}) = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}.$$

the more N we observe, the closer $\bar{x}$ is to μ

Extra: CDF example

Say we have a normal:

probability density function, sometimes notated as $f(x)$.

PDF

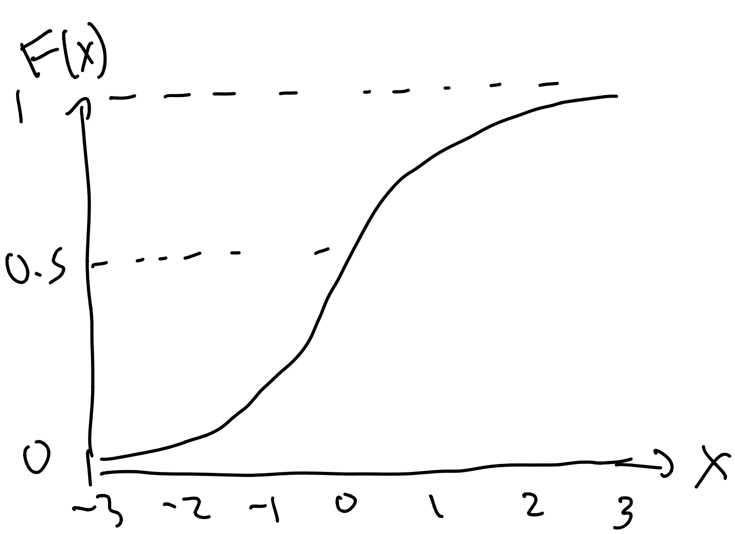
$\mathbb{P}(X \leq -1)$ is the integral from $-\infty$ to -1 of $f(x)$

A CDF is an integral from $-\infty$ to (some place) of $f(x)$

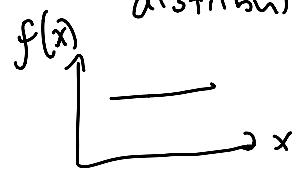
$$CDF(x) = \int_{-\infty}^{x} f(x')\,dx',$$ often notated as $F(x)$.

So $\mathbb{P}(X \leq -1) = CDF(-1)$

The further right we integrate to, we can only <u>add</u> to the integral, so $CDF(x)$ cannot decrease w/ x.



Q: draw the CDF of a uniform distribution



(Hint: integrate up to x)