Introduction to frequentist inference

Lecture 10a (STAT 24400 F24)

Framework for frequentist inference

Setup

- We are interested in learning about a parameter θ (e.g. $\lambda, \alpha, \beta, \mu, \sigma^2, \cdots$).
  which is a fixed, but unknown, constant.

  (Compare to Bayesian stats: θ is random, with a prior distribution)

- We observe data X whose distribution is *parametrized* by θ.

Based on the observed X, we might ask:

- **Parameter estimation**
  Which value of θ is the most plausible in light of the data?
- **Confidence intervals**
  Which values of θ are plausible in light of the data?
- **Hypothesis testing**
  For a particular value of θ, is this value plausible in light of the data?

  (Bayesian stats: instead work with the posterior distribution)

Example: calibrating the X-ray

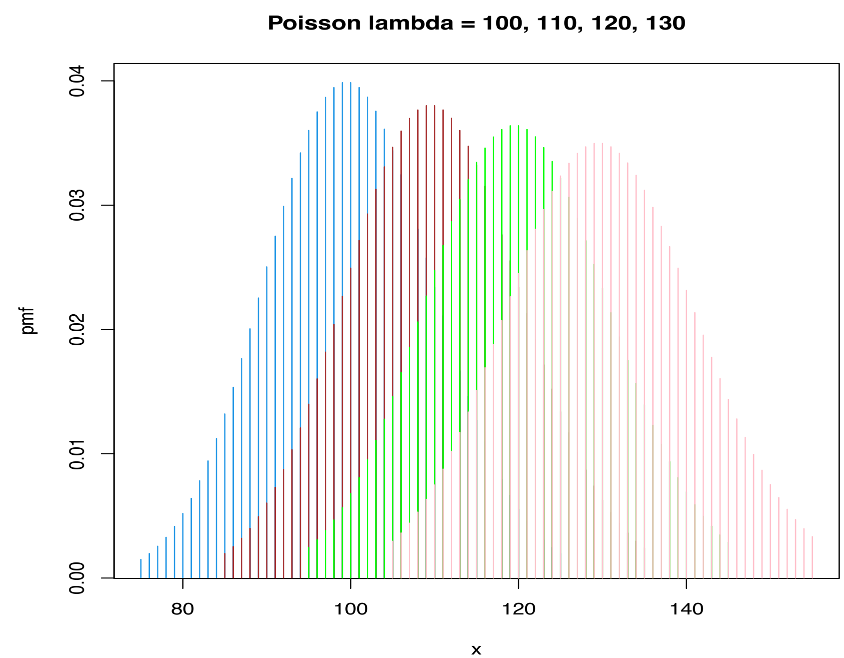
We have an X-ray beam whose output follows a Poisson(λ) distribution.

The intensity parameter λ can be set to 100, 110, 120, or 130.
(Since λ is unknown, we should assume we cannot see the setting.)

If $\lambda = \ldots$	then likely X values are ...	
100	84–117	
110	93–128	$\leftarrow$ each range contains 90% probability
120	102–138	
130	112–149	

If we observe $X = 108$,
we might say $\lambda = 100, 110, 120$ are plausible (and $\lambda = 130$ is not)

Q: Which value should we pick? How do we quantify our errors?



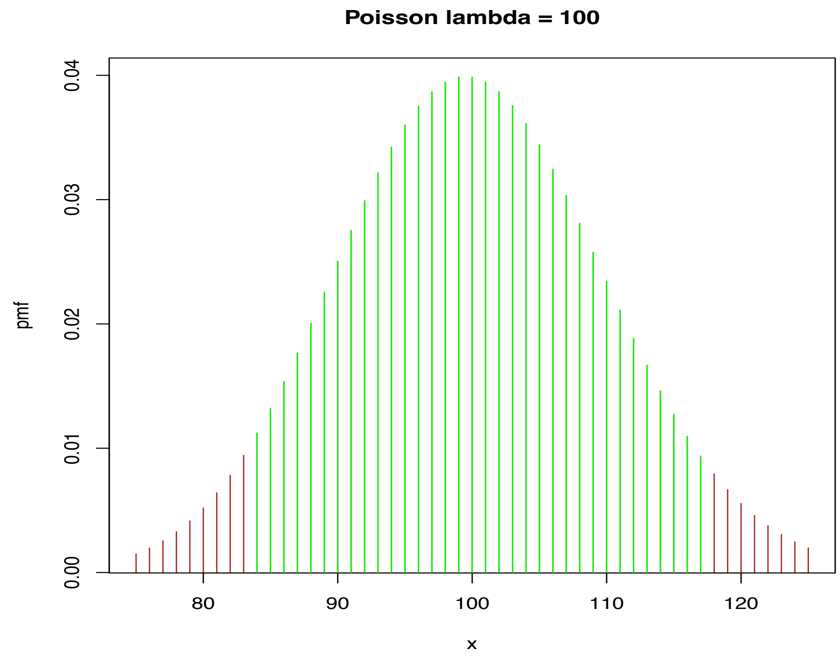
Example: calibrating the X-ray (testing a hypothesis)

Suppose person A claims that the X-ray is set at $\lambda = 100$,
and person B wants to test if this claim is true.

Here is person B's reasoning:

- If the X-ray scanner were really set to $\lambda = 100$,
  then the likely values for X are 84–117 (with a 90% chance)
- So if we run the scanner and measure a value X outside this range,
  we can conclude that $\lambda \neq 100$, because $\lambda = 100$ is *not* plausible

Our error rate (for this hypothesis test) is 10%:
if $\lambda = 100$, we have a 10% chance of concluding $\lambda \neq 100$



Caution: incorrect interpretations

Caution An **incorrect** interpretation:

"If we observe X outside the range,
then there's only a 10% chance that $\lambda = 100$ actually is true"

- We aren't in a Bayesian framework — $\mathbb{1}_{\lambda=100}$ isn't random
- And even in a Bayesian framework,
  $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$ so our calculation is likely wrong

$\mathbb{P}(X \text{ outside range} \mid \lambda = 100)$ $\mathbb{P}(\lambda = 100 \mid X \text{ outside range})$

Intro to **confidence intervals**

- We are interested in learning about a parameter θ,
  which is a fixed, but unknown, constant.
- We observe data X whose distribution is *parametrized* by θ

Goal of confidence interval:

Find a range for plausible values of θ, after observing data.

After observing X, we return a range

$$[\ \text{lowerbound}(X)\ ,\ \text{upperbound}(X)\ ]$$

such that for θ,

the unknown true parameter giving the distribution of the data

$$\mathbb{P}\big(\ \text{lowerbound}(X)\ \leq\ \theta\ \leq\ \text{upperbound}(X)\ \big) \geq 1 - \alpha$$

probability over $X \sim$ (distrib. with parameter θ) tolerate error $\leq \alpha$

Note: This doesn't contradict our framework where θ is not random,
because the endpoints of the inequality are random (depend on X).

From data to confidence intervals (example to illustrate the idea)

For the X-ray scanner, if we observe X, we would build a range around X, e.g.,

$$[\text{ lowerbound}(X), \text{ upperbound}(X) \ ] = [X - 20, X + 20]$$

We think it's plausible that the true value of λ is in this interval.

X, the random variable, has high probability to occur near the true λ:

- If $\lambda = 100$, $\mathbb{P}(X - 20 \leq \lambda \leq X + 20) = \mathbb{P}(|\lambda - X| \leq 20) = 0.960$
- If $\lambda = 110$, $\mathbb{P}(X - 20 \leq \lambda \leq X + 20) = \mathbb{P}(|\lambda - X| \leq 20) = 0.950$
- If $\lambda = 120$, $\mathbb{P}(X - 20 \leq \lambda \leq X + 20) = \mathbb{P}(|\lambda - X| \leq 20) = 0.939$
- If $\lambda = 130$, $\mathbb{P}(X - 20 \leq \lambda \leq X + 20) = \mathbb{P}(|\lambda - X| \leq 20) = 0.928$

So, if $100, 110, 120, 130$ are the only possible values of λ, then this (random) interval $[X - 20, X + 20]$ is a 92.8% confidence interval for λ, it has a 92.8% coverage.

Questions in confidence intervals

Further questions:

- How should we construct the range [lowerbound(X), upperbound(X)]?

  That is how do we choose the lowerbound(X) and upperbound(X), functions of data, such that the confidence interval constructed is

  - not too wide
    — so that the range is still informative and practically useful
  - not too narrow
    — so the range still has good coverage to contain the true parameter

- How does this method relate to Bayesian inference?

Intro to **hypothesis testing**

- We are interested in learning about a parameter θ,
  which is a fixed, but unknown, constant
- We observe data X whose distribution is *parametrized* by θ

Goal of hypothesis testing:

We are interested in testing a *null hypothesis* $\theta = \theta_0$,
versus an *alternative hypothesis* (e.g., $\theta = \theta_1$ or $\theta \neq \theta_0$ or $\theta > \theta_0$)

Generally, our goal is to disprove the null.

e.g., null = "vaccine has zero effect on risk of catching the disease",
alt. = "vaccine reduces risk"

We construct a range of likely X values, assuming $\theta = \theta_0$ is true, such that

$$\mathbb{P}\big(X \in (\text{the range of values})\big) \geq 1 - \alpha$$

probability over $X \sim$ (distrib. with parameter θ_0)

From data to hypothesis testing (decision and error rate)

After observing X, we return a decision:
If X falls into the plausible range, we *do not reject* the null.
If X falls outside the plausible range, we *reject* the null (& conclude $\theta \neq \theta_0$)

Then we have a $\leq \alpha$ chance of incorrectly rejecting a true null.

Further questions:

- What's our chance of *failing* to reject, if the alternative is true?
- How should we construct the range of X values?
- How does hypothesis testing relate to p-values (coming soon)?
- How does hypothesis testing relate to confidence intervals?
- What is the problem of *multiple testing*?