Advanced Probability Questions

Joyce

Probability Questions

How are probability interview questions asked?

- Asked directly
  - Combination / Permutation
  - Mean / Variance / Relationship of distributions
  - Coin toss questions
- Asked with real world examples
  - Need to figure out which distribution / probability rules to use
- Asked in Hypothesis testing problems

Frequently Asked Questions

- Combinations / Permutations
- Bernoulli / Binomial (coin toss, conversion rate, etc)
- Normal distribution (CLT, mostly with hypothesis testing / power analysis)
- Bayes Rules
- Poisson, Geometric
- More advanced:
  - Truncated Normal
  - Multivariate Normal
  - Zero-inflated Poisson

Example 1 - 1

A chopstick factory want to know the mean length of its products. They had a group of samples and a technician measure and log the length of each sample. However, he made a mistake of logging data as ‘NA’ if the sample’s length > 32 cm. What is the estimated population mean of the products’ length?

Assume we know total number of samples and number of ‘NA’ values

Example 1 - continued

What assumptions you need to make?

- Each product’s length $i.i.d \sim N(\mu, \sigma)$

What is the parameter we want to estimate?

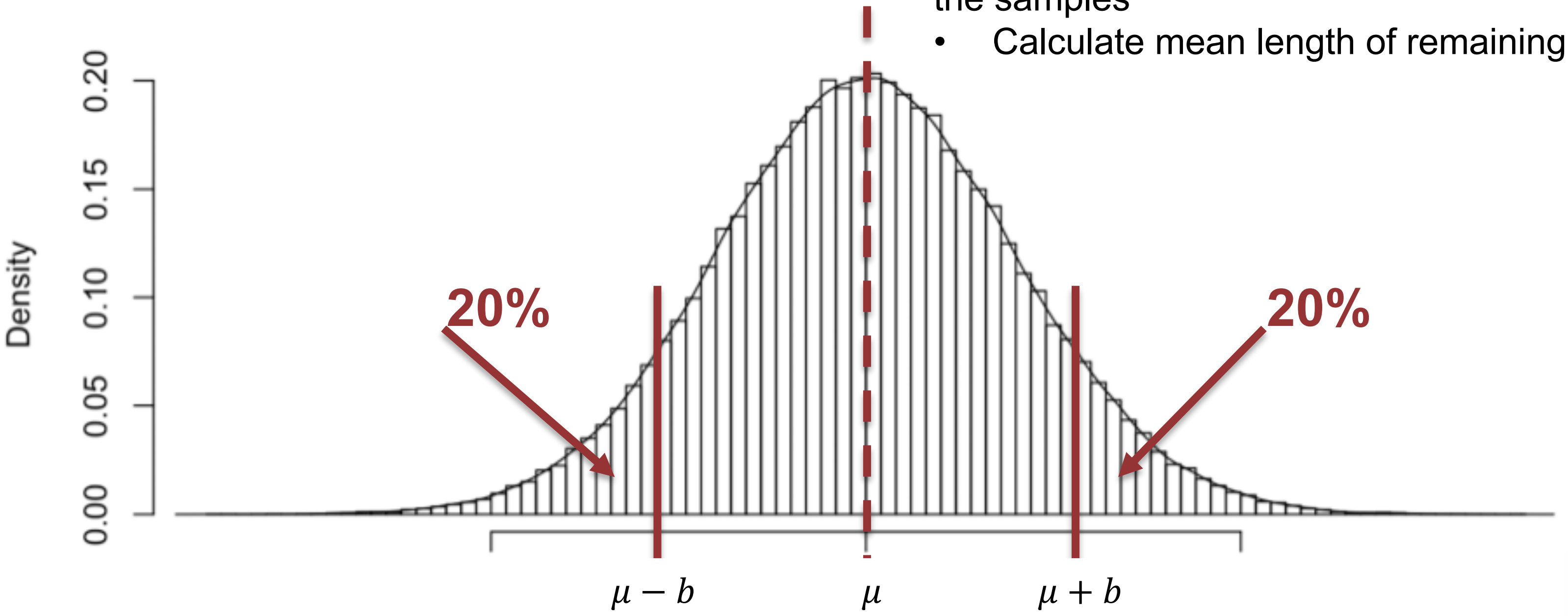
- μ

Symmetric Quantiles of Normal

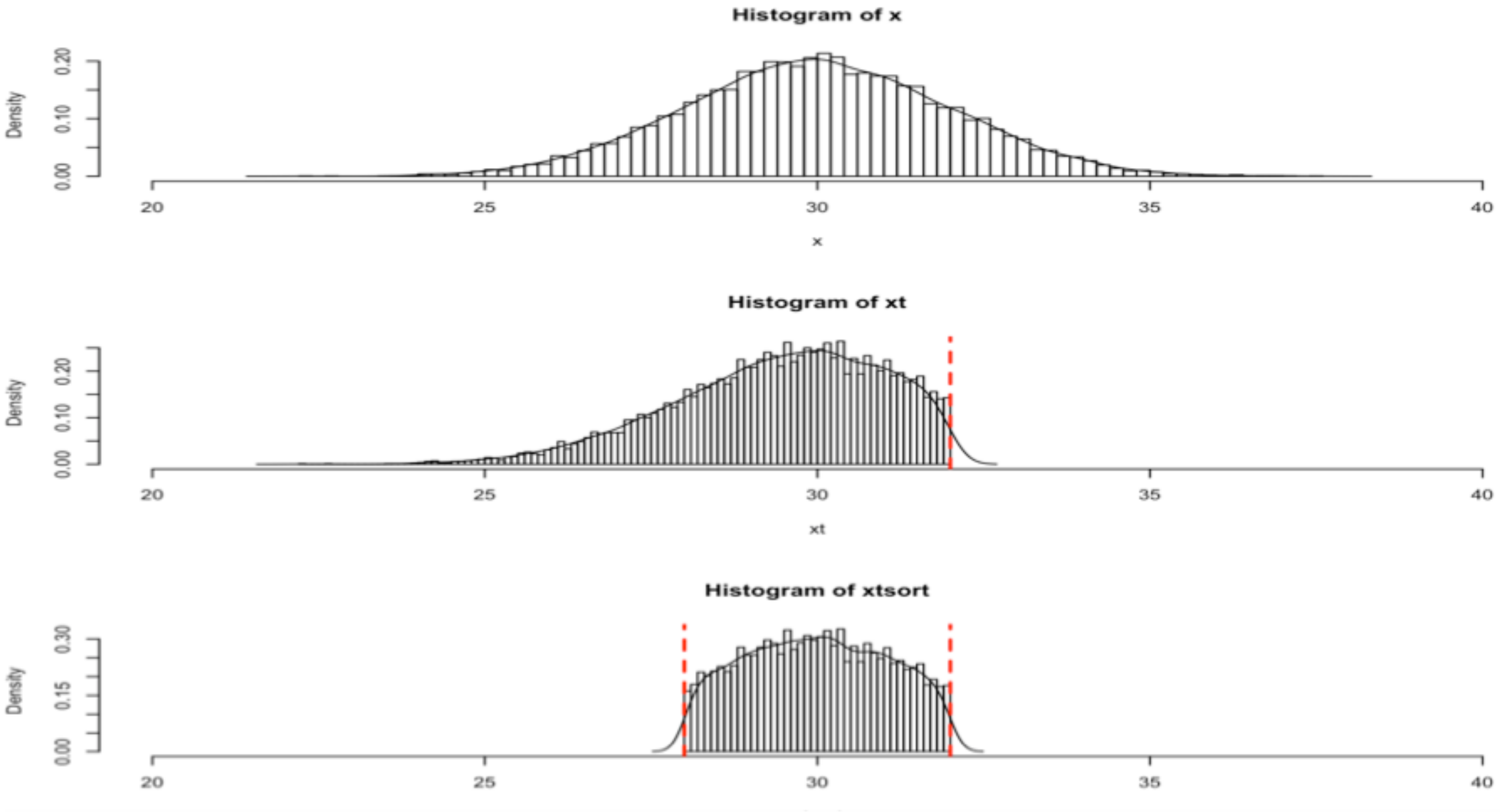
Example 1000 samples, 200 are ‘NA’

Solution

- Sort the samples
- Remove bottom 20% and ‘NA’ of the samples
- Calculate mean length of remaining samples



Solution – R simulation



Example 1 - 2

What if the technician only recorded non-NA values?

Do **NOT** know total number of samples and number of 'NA's

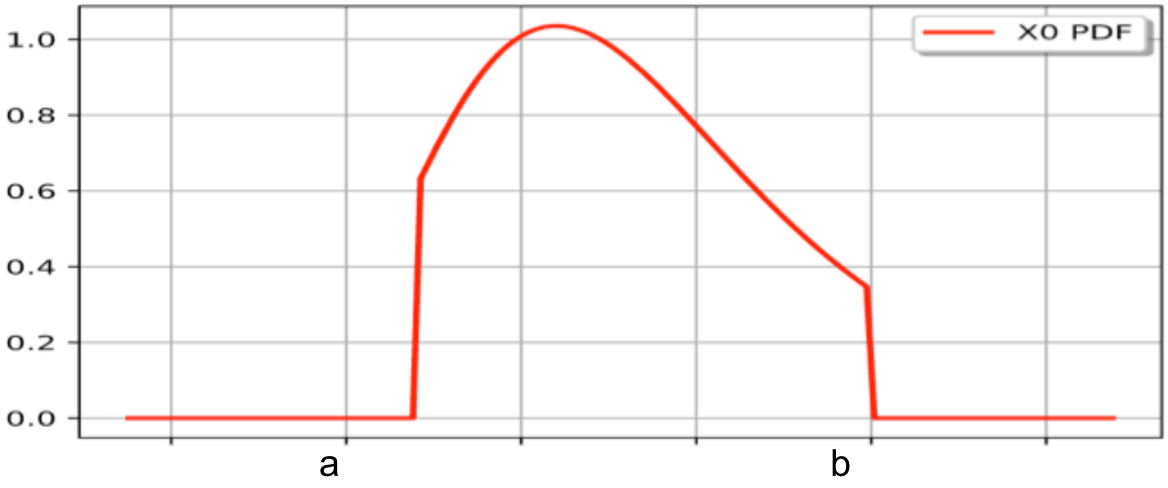
We need to estimate from the sample's distribution

- Truncated Normal Distribution
- Conditional Probability
- Likelihood Function
- MLE

Truncated Normal Distribution

the **truncated normal distribution** is the probability distribution derived from that of a normal distributed random variable by bounding the random variable from either below or above (or both)

$$x \sim N(\mu, \sigma), x \in (a, b)$$



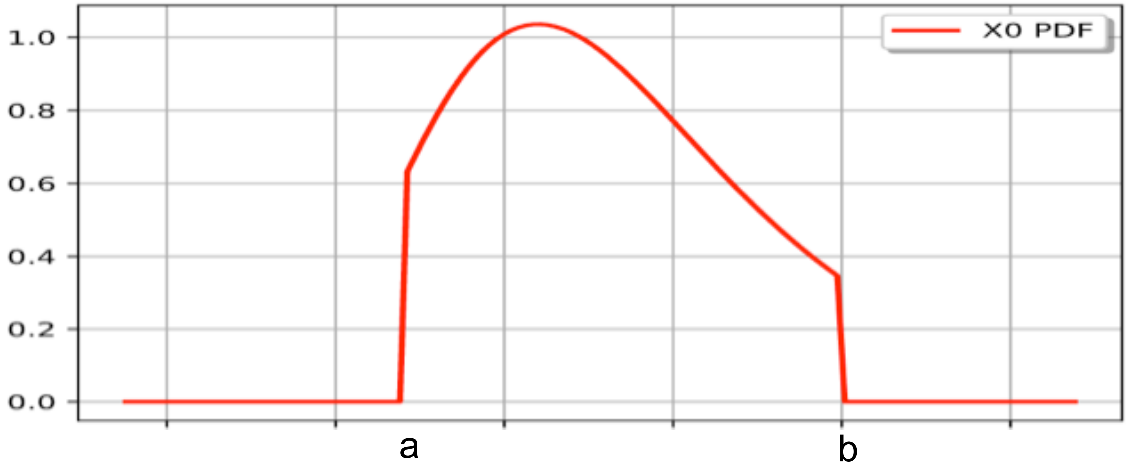
https://en.wikipedia.org/wiki/Truncated_normal_distribution

Density Function of Truncated Normal

For regular normal distribution $N(\mu,\sigma)$, $f(x) = p.d.f\ of\ N(\mu,\sigma)$, $\Phi(x) = c.d.f\ of\ N(\mu,\sigma)$

For truncated normal distribution $NT(\mu,\sigma)$,

$$f_{nt}(x) = f(x|a < x < b) = \frac{f(x)}{f(x\sim N(\mu,\sigma)\ \&\ a < x < b)} = \begin{cases} 0 & ,x < a, x > b \\ \dfrac{f(x)}{\Phi(b) - \Phi(a)} & ,a < x < b \end{cases}$$



https://en.wikipedia.org/wiki/Truncated_normal_distribution

Likelihood Function & MLE

a **likelihood function** is a **function** of the parameters of a statistical model given data

Density function: function of data given parameters $f(x|\theta)$
Likelihood function: function of parameters given data $\mathrm{L}(\theta|x)$

$$L(\theta|x) = f(x|\theta)$$

When you have data observations and you want to estimate θ, you want the most 'likely' estimation, which is to maximize the likelihood function. This estimate is called **MLE** (maximum likelihood estimator)

$$\hat{\theta} \in \{\arg\max L(\theta|x)\}$$

How to calculate: take log of likelihood function, take derivatives

Most commonly used estimator!! Important!!

https://en.wikipedia.org/wiki/Likelihood_function

1-2 Solution

Calculate the MLE of μ using likelihood function of $L(x|)$ $\mathrm{L}(\mathrm{x}|\mu,\sigma) = f_{nt}(x|\mu,\sigma)$

Closed form solution from Wikipedia, Don't need to remember this!

Let $\alpha = (a - \mu)/\sigma$ and $\beta = (b - \mu)/\sigma$

$$\mathrm{E}(X \mid a < X < b) = \mu + \sigma \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} = \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}$$

Test example in R

https://en.wikipedia.org/wiki/Truncated_normal_distribution

Example 1 Summary

If quantile is known, use **Symmetric Quantiles** property of Normal distribution

If quantile not known, estimate with **MLE** of **Truncated Normal** distribution

- Truncated Normal distribution
- Conditional Probability
- Likelihood Function
- MLE

Example 2

A user on your website will send a signal for every second the user is logged in. Your logging system will open a new file when a user log in, write down each signal received, close the file when the user log-out. However, the system is experiencing some problems which randomly fail to log a signal. The probability of failure is consistent.

You got a user’s file showed the last signal is the 1000th second. What is the estimated time spend for this login?

Questions

What are we trying to estimate?

User's actual active time = logged time + consecutive failed logging after last log
E(actual active time) = 1000 + E(consecutive failed logging)

What is the probability of one failed logging ?

Bernoulli distribution, what is the p??

What is the probability of n consecutive failed logging ?

Geometric distribution

Geometric Distribution

The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set { 1, 2, 3, ...}

$$E(X) = \frac{1}{p}$$

The probability distribution of the number Y = X − 1 of failures before the first success, supported on the set { 0, 1, 2, 3, ... }

$$E(Y) = \frac{1}{P} - 1$$

Solution: $1000 + \frac{1}{\hat{p}} - 1,\ \hat{p} = \frac{\#\ of\ missed\ loggings}{1000}$

Learning:

- Abstract probability distributions from real problem
- Think about how to estimate parameter
- Geometric Distribution

Example 3

We are testing a new version of website to users, every day we select 1% users to see new version. What is the expected waiting time for a user to see the new version?

Example 4

You have a 0.1% chance of picking up a coin with both heads and a 99.9% chance that you pick up a fair coin. You picked your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?

Bayes Rule & Law of Total Probability

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$\Pr(A) = \sum_n \Pr(A \mid B_n)\Pr(B_n)$$

You have a 10% chance of picking up a coin with both heads and a 90% chance that you pick up a fair coin. You picked your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?

$$P(picked\ fair\ coin \mid 10\ heads) = P(10\ heads\ | picked\ fair\ coin) * \frac{P(fair\ coin)}{P(10\ heads)}$$

$$= 0.5^{10} * \frac{0.1}{P(10\ heads\ |fair) * P(fair) + P(10\ heads|unfair) * P(unfair)}$$

$$= \frac{0.5^{10} * 0.1}{0.5^{10} * 0.9 + 1 * 0.1}$$