AQP Math Foundations: Random Variables, Estimators, and Accuracy Bounds

Chris Jermaine

Rice University

Sampling Basics

Consider the following data set, or "population":

$$\langle 3, 4, 5, 6, 9, 10, 12, 13, 15, 19 \rangle.$$

We want to estimate:

```
SELECT SUM (R.a)
FROM R
```

Sampling Basics

To do this, first draw a "sample":

- Simulate rolling a 10-sided die $n = 5$ times (via a PRNG)
- If obtain a j on trial i, then ith item in sample is jth value
- Imagine I roll $\langle 6,3,5,3,9 \rangle$
- Then associated sample of `R.a` values is $\langle 10,5,9,5,15 \rangle$
- This is "simple random sampling with replacement"

Sampling Basics

Then compute query result on sample

- This gives us 44
- Then scale up by a factor of 2
- Compensates for the fact our sample size is $5/10 = 1/2$ of the values
- This gives us 88 (real answer is 96)

Random variables

This is the intuitive view...

But how good is this estimate?

Now we need some math

Denote by:

- t_i the value of the ith item in the population
- X_i the random variable controlling the number of times t_i appears in the sample
- (What's a random variable?)

Example:

- Original data: $\langle 3,4,5,6,9,10,12,13,15,19 \rangle$
- Sample of `R.a` values is $\langle 10,5,9,5,15 \rangle$
- $X_6 = 1$ and $X_3 = 2$... why?
- The sixth item ($t_6 = 10$) appears once
- The third item ($t_3 = 5$) appears twice

Our Estimate as a RV

We can write our sampling-based estimator as a RV:

$$Y = 2\sum_{i=1}^{N} X_i t_i = 2 \sum_{i \in \text{sample}} t_i$$

with our sample,

$$\hat{Y} = 88$$

where:

- $N = 10$ is the population size
- "sample" denotes the set of distinct items in the sample
- $\hat{Y}$ is the observed value for Y

Expectation of Random variables

"Goodness" will generally be defined using the "expectation" of our estimator

Where our "estimator" is the RV Y

Generally, we want $E[Y] \approx Q$...

That is, we like our estimate to be correct on expectation

Or "unbiased"

Expectation of Random variables

The "expected value" of a RV:

- The average value of an infinite number of trials over the RV

In general:

- $E[X] = x_1 Pr[X = x_1] + x_2 Pr[X = x_2] + x_3 Pr[X = x_3] + \ldots$
- Or, $E[X] = \int_{-\infty}^{\infty} x f(x) dx$ if continuous

In our case, defined as:

- $E[X_i] = \sum_{j=0}^{5} j \times Pr[X_i = j] = \frac{1}{2}$
- Note that each X_i is binomial
- Makes sense, as we select $\frac{1}{2}$ of the items

Estimators and Bias

Since:

- $E[X+Y] = E[X] + E[Y]$ (expectation is linear)
- $E[cX] = cE[X]$ (here, c is a constant)

We have:

$$E[Y] = E[2\sum_{i=1}^{N} X_i t_i] = E\left[\sum_{i=1}^{N} \frac{X_i t_i}{E[X_i]}\right] = \sum_{i=1}^{N} E\left[\frac{X_i t_i}{E[X_i]}\right]$$

$$= \sum_{i=1}^{N} \frac{E[X_i] t_i}{E[X_i]} = \sum_{i=1}^{N} t_i = Q \text{ (query result)}$$

Estimators and Bias

Note that $E[Y] = Q$

Hence the estimator is "unbiased"

- This means it is correct on expectation
- Bias is one of two key types of estimation error...
- so unbiased is good!

Accuracy

Unbiased is good...

But not the whole story

We also worry about variance, or variability of the estimate

How to compute variance?

- Will spend a lot of time on this question...
- But one classical approach is the "Central Limit Theorem" (CLT)
- Applies when each sample is i.i.d. (SRSWR is only standard i.i.d. scheme)

Central Limit Theorem

To apply the CLT:

- First compute the sample variance of sampled items; $17.2 =$

$$\frac{1}{4}\left((10-8.8)^2+(5-8.8)^2+(9-8.8)^2+(5-8.8)^2+(15-8.8)^2\right)$$

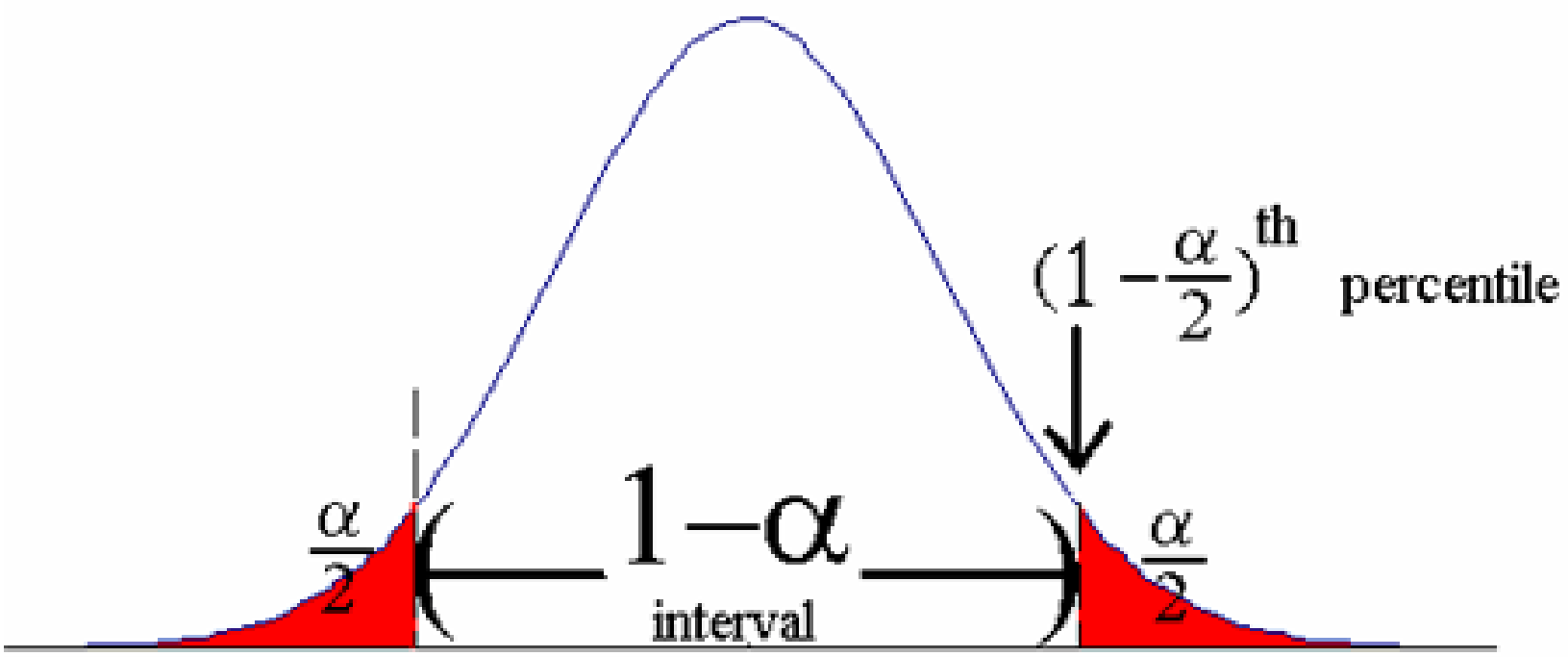
- Note, 8.8 is the average of the five
- CLT says $\frac{1}{n}\sum_{i=1}^{N} X_i t_i$ is Normal for large sample (n is sample size, N population size)
- CLT also says if σ^2 is population variance (variance of t_i's), variance of $\frac{1}{n}\sum_{i=1}^{N} X_i t_i$ is σ^2/n
- So variance of $Y/n = \frac{2}{n}\sum_{i=1}^{N} X_i t_i$ is $2^2\sigma^2/n$

CLT-based variance

So what's the variance of Y?

- Variance of $(Y/n)n$ is $(2^2\sigma^2/n)5^2$ or $10^2(17.2/5) = 344$
- Note: only an approximation since 17.2 approximates population variance

How does this tell us how accurate we are?



- Since $\pm 2\sigma$ contains approximately 95% of a Normal's mass...
- And we're unbiased
- Means we can guess answer is $88 \pm 2(344)^{\frac{1}{2}} = 88 \pm 37$

A Math Model for Sampling

CLT doesn't hold in every case

Requires i.i.d.

Can we develop tools that work with other, non i.i.d. schemes?

Yes!

A Math Model for Sampling

In general, let a "sample" be a list of the form:

$$\langle (t_1, X_1), (t_2, X_2), \ldots, (t_n, X_n) \rangle$$

- t_i is ith tuple, X_i number of times it appears in sample
- Example: in Bernoulli ("coin-flip") sampling, $X_i = 1$ if obtain "heads" on flip i
- Example: in SRSWR, X_i is the number of times i selected

An estimator Y is a function $\mathscr{F}$ applied to this list

$$Y = \mathscr{F}\left(\langle (t_1, X_1), (t_2, X_2), \ldots, (t_n, X_n) \rangle\right)$$

A Math Model for Sampling

- In general,

$$\mathscr{F}\left(\langle \ldots, (t_{j-1}, X_{j-1}), (t_j, 0), (t_{j+1}, X_{j+1}), \ldots \rangle\right) = \mathscr{F}\left(\langle \ldots, (t_{j-1}, X_{j-1}), (t_{j+1}, X_{j+1}), \ldots \rangle\right)$$

- That is, $\mathscr{F}$ cannot "look at" a non-sampled item

A Math Model for Sampling

- Accuracy generally has two parts
- (1) Bias:

$$bias(Y) = E[Y] - Q$$

- (2) Variance:

$$\sigma^2(Y) = E[(Y - E[Y])^2] = E[Y^2] - E^2[Y]$$

- Bias + variance is small means high-quality

Example

Consider the query:

```
SELECT SUM(r.extendedprice * (1.0 - r.tax))
FROM R as r
WHERE l.suppkey = 1234
```

- Set t_j to be `r.extendedprice * (1.0 - r.tax)` if `r.suppkey = 1234`
- Where `r` is the jth tuple in `R`
- Otherwise, if `r.suppkey <> 1234` then $t_j = 0$
- Then:

$$Q = \sum_j t_j$$

Example

Imagine our data set is

$$\langle t_1, t_2, \ldots, t_{10} \rangle = \langle 3, 4, \ldots, 19 \rangle$$

Take a sample (with replacement) of size 5... then let

$$Y = \mathscr{F}\left(\langle (t_1, X_1), \ldots, (t_{10}, X_{10}) \rangle\right) = \frac{10}{5} \sum_{j=1}^{10} t_j X_j$$

And

$$E[Y] = E[\frac{10}{5} \sum_{j=1}^{10} t_j X_j] = 2 \sum_{j=1}^{10} t_j E[X_j] = 2 \sum_{j=1}^{10} t_j \frac{1}{2} = Q$$

So our estimate is unbasied

What about the variance?

$$\sigma^2(Y) = E[(Y - E[Y])^2] = E[Y^2] - E^2[Y]$$

Since $E[Y] = Q$, we have $E^2[Y] = Q^2$

What about $E[Y^2]$?

- A little more involved:

$$E[Y^2] = E\left[\left(2\sum_j t_j X_j)\right)^2\right] = E\left[4\sum_i\sum_j t_i t_j X_i X_j\right]$$

$$= E\left[4\sum_i t_i^2 X_i^2 + 8\sum_{i<j} t_i t_j X_i X_j\right] = 4\sum_i (t_i^2 E[X_i^2]) + 8\sum_{i<j} t_i t_j E[X_i X_j]$$

- We now have two summations...

Variance

We now have two summations...

- One with $E[X_i^2]$... 2nd moment of binomial, or $np + n(n-1)p^2 = \frac{1}{2} + \frac{1}{5} = \frac{7}{10}$
- $E[X_i X_j]$ is more non-standard, but is $n(n-1)p^2 = \frac{1}{5}$
- (note: why is $E[X_i X_j] \neq E[X_i][X_j]$?)
- Plugging this into the above equation, we have:

$$E[Y^2] = \frac{14}{5}\sum_i t_i^2 + \frac{8}{5}\sum_{i<j} t_i t_j$$

Variance

And so the variance of Y, denoted as $\sigma^2(Y)$, is:

$$\sigma^2(Y) = \frac{14}{5}\sum_i t_i^2 + \frac{8}{5}\sum_{i<j} t_i t_j - Q^2$$

Plugging in the actual values for our dataset, we have:

$$\sigma^2(Y) = \frac{14}{5} \times 1166 + \frac{8}{5} \times 4025 - 9216 = 488.8$$

- 488.8 differs from $17.2 \times \frac{100}{5} = 344$ via the CLT
- Why?

Horvitz-Thompson Estimators

Most common sampling scheme:

- Each item appears at most once ($X_i = 0$ or 1)
- Where π_i is probability that item i in the sample
- So $E[X_i] = \pi_i$
- Often, $\pi_i = \pi_j$ forall i, j (SRSWOR, for example)

Then the "Horvitz-Thompson" estimator of the sum is

$$Y = \sum_{i \in \text{sample}} \frac{t_i}{\pi_i}.$$

- Always unbiased
- Can be very accurate by giving high sampling probability to "important" items

Variance of HT

- For a HT estimator Y...
- Simple algebra gives us:

$$\sigma^2(Y) = E[Y^2] - E^2[Y] = E\left[\left(\sum_i \frac{X_i t_i}{\pi_i}\right)^2\right] - \left(\sum_i t_i\right)^2$$

$$= E\left[\sum_i \sum_j \frac{X_i t_i}{\pi_i} \frac{X_j t_j}{\pi_j}\right] - \sum_i \sum_j t_i t_j$$

$$= \sum_i \sum_j \frac{\pi_{ij} t_i t_j}{\pi_i \pi_j} - \sum_i \sum_j t_i t_j = \sum_i \sum_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) t_i t_j.$$

- Note that π_{ij} is the probability of BOTH X_i and X_j being one
- Depends on sampling scheme...

Estimating Variance of HT

In practice, need to estimate $\sigma^2(Y)$ from the sample

- Note that $\sigma^2(Y)$ is a sum
- Over the cross-product of the dataset with itself
- So we can derive a HT estimator for $\sigma^2(Y)$:

$$\hat{\sigma}^2(Y) = \sum_i \sum_j \frac{X_i X_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) t_i t_j = \sum_{i,j \in \text{sample}} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) t_i t_j.$$

Example: Estimating Variance of HT

Consider SRSWOR

We have the following data set:

$$\langle 3,4,5,6,9,10,12,13,15,19\rangle.$$

We want to estimate:

```
SELECT SUM (R.a)
FROM R
```

And we sample $\langle 3,5,6,15,19\rangle$

- HT estimator for sum is $2(3+5+6+15+19) = 96$
- What about variance?

Example: Estimating Variance of HT

We previously had

$$\hat{\sigma}^2(Y) = \sum_{i,j \in \text{sample}} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) t_i t_j.$$

What is π_{ij}?

- Two cases: $i = j$, $i \neq j$
- If $i \neq j$, $\pi_{ij} = \frac{5}{10}\frac{4}{9} = \frac{2}{9}$
- If $i = j$, $\pi_{ij} = \frac{5}{10} = \frac{1}{2}$

From Bias and Variance To Accuracy

We've now extensively discussed bias and variance

How to translate to accuracy?

- Typically, we want a "confidence bound"
- Probabilistic guarantee of the form...
- "There is a $p \times 100\%$ chance that the true answer to the query is within the range l to h."

Many ways to come up with confidence bounds...

Central Limit Theorem

Technically, CLT applies only to i.i.d. case in limit

But often, we find that errors are normally distributed in practice

- So we assume that Y can be modeled as:

$$Y \approx Q + \mathcal{N}(0, \sigma^2(Y))$$

- Then choose numbers lo and hi so that:

$$p = \int_{\text{lo}}^{\text{hi}} f_{\mathcal{N}}(x)dx$$

- where $f_{\mathcal{N}}$ is the probability density function of $\mathcal{N}$
- Then, there is a $p \times 100\%$ chance that $\text{lo} \leq Q - Y \leq \text{hi}$

Chebyshev Bounds

If CLT makes one uncomfortable, use distribution free bounds

- Chebyshev's inequality implies for unbiased estimate Y,

$$Pr[|Y - Q| \geq p^{-\frac{1}{2}}\sigma(Y)] \leq p$$

- So there is a $p \times 100\%$ chance that Q is between $Y - p^{-\frac{1}{2}}\sigma(Y)$ and $Y + p^{-\frac{1}{2}}\sigma(Y)$
- Much looser than CLT bounds!
- In our initial example, bounds go from 88 ± 37 to 88 ± 117.32

Hoeffding Bounds

Some bounds don't even require variance

- Hoeffding bounds applicable when $Y = \frac{1}{n}\sum_i X_i$
- and where the value of X_i ranges from low_i to hi_i
- Then,

$$Pr[|Y - E[Y]| \geq d] \leq 2\exp(-\frac{2d^2n^2}{\sum_i(hi_i - low_i)^2})$$

- But MUCH looser!

Hoeffding Bounds

Recall we sampled $\langle 10, 5, 9, 5, 15 \rangle$

- Multiply by 10 to get $\langle 100, 50, 90, 50, 150 \rangle$
- Mean of this sequence is unbiased for Q
- Assume 50 and 150 bound numbers to sample
- Then approximate $\sum_i (hi_i - low_i)^2$ by $n(150-50)^2 = 50,000$
- Now, solve for d:

$$0.05 = 2\exp(-\frac{2d^2n^2}{50,000})$$

- Gives $d = 192.06$
- 88 ± 192.06 is a 95% confidence interval
- Mostly useless!!

What about bias?

Not all estimators unbiased

- Consider:

```
SELECT COUNT (*)
FROM R as r1, R as r2
WHERE r1.a BETWEEN r2.a - 3 AND r2.a + 3
```

- We draw a size $n = 5$ with-replacement sample of `R`
- Join the sample with itself
- Scale result by $\frac{1}{n^2\pi^2} = 4$
- Will be biased

What about bias?

Why biased? Begin with $E[Y]$

$$E[Y] = E\left[4\sum_{j,k} I(t_j.\mathtt{a} \ \texttt{BETWEEN}\ t_k.\mathtt{a}-3 \ \texttt{AND}\ t_k.\mathtt{a}+3)X_jX_k\right]$$

- I is the indicator function
- Now use $I(t_j.\mathtt{a}, t_k.\mathtt{a})$ as shorthand for $I(t_j.\mathtt{a}$ BETWEEN $t_k.\mathtt{a}-3$ AND $t_k.\mathtt{a}+3)$

$$\begin{aligned} E[Y] &= E\left[4\sum_j\sum_k I(t_j.\mathtt{a}, t_k.\mathtt{a})\right] \\ &= E\left[4\sum_j I(t_j.\mathtt{a}, t_j.\mathtt{a})X_j^2 + 4\times 2\sum_{j<k} I(t_j.\mathtt{a}, t_k.\mathtt{a})X_jX_k\right] \\ &= 4\sum_j I(t_j.\mathtt{a}, t_j.\mathtt{a})E[X_j^2] + 4\times 2\sum_{j<k} I(t_j.\mathtt{a}, t_k.\mathtt{a})E[X_jX_k] \end{aligned}$$

What about bias?

But not correct on expectation

- Recall that $E[X_j^2] = \frac{7}{10}$, $E[X_j X_k] = \frac{1}{5}$
- So we have:

$$E[Y] = \frac{14}{5}\sum_j I(t_j.\mathsf{a}, t_j.\mathsf{a}) + \frac{8}{5}\sum_{j<k} I(t_j.\mathsf{a}, t_k.\mathsf{a})$$

- But $Q = \sum_j I(t_j.\mathsf{a}, t_j.\mathsf{a}) + 2\sum_{j<k} I(t_j.\mathsf{a}, t_k.\mathsf{a})$
- So $Q \neq E[Y]$

How to handle bias?

Three ways:

- (1) Ignore it!
  - Many estimators are biased
  - But asymptotically unbiased

How to handle bias?

Three ways:

- (2) Estimate and counter-act
  - Useful if bias large
  - But can result in large standard error $((bias^2(Y) + \sigma^2(Y))^{1/2})$
  - If estimate for bias has high variance

How to handle bias?

Three ways:

- (3) Estimate bias and use in bounds
  - Rule-of-thumb is that for $p \leq 0.95$...
  - Generally safe to replace std deviation with std error in bounds
  - As long as the ratio of $bias(Y)$ to $\sigma(Y) < 0.5$

That's it for math foundations

Questions?