Note 7

Review

Add pop-up quiz material from paper

Query Averaging

- Hides variance of systems preformance
- Can be very good in some queries, bad in others

Significance test

- Is our result due to our particular test sets or is it a fair indication of how our system preforms?
- Variance across systems and within systems. Want to make sure result is not coming from your dataset

Background

- **p-value** probability of obtaining data as extreme was observed, if null hypothesis was true (e.g, if observation is totally random)
- null hypotheis =observation @ random. if p is smaller than confident leve, we say null is false.
- We seek to reject null hypothesis. Observation is random result. Small p-val is good.
- assume something, try to reject it

Stat review

- Distribution is about some random variable
- In IR what is our random variable? Our query
- What we return from some query is random
- We assume some hypothesis. one case after another. Then we try to reject

Sign Test

- NH: diff median is zero between samples from two continuous distributions

Wilcoxon Signed rank test

- same sign from sign test. then rank them by absolute rank
- sum of + is some val , sum of - is some val
- take difference. should be as close to 0 as possible or less that critical value.
- If its not you cannot reject NH
- NH: Data are paired and come from same population

- Don't believe it gaussian. Don't assume any distribution. Non-parametric
- Don't care value, you care the order. Is this system better or not?

Paired t-test

- NH: Difference between two responses measured on same statistical unit has a zero mean val.
- Strong assumption of equal variance between your systems

One tail

- one system is better than other. two-tail , diverse outcomes between systems

Where do we get relevance?

- Old school : Human annotation. 1) Enter query into my system. And rank results 2) annotator has to guess underlying information need
- Pooling: Pool results from other search engines. Top k results
- Can never be exhaustive

Kappa statistic

- Two annotaters are consistent with each other
- P(A) - P(E) / (1-P(E))
- Denom is primarily for normalization
- Want as large as possible (largest val is 1)
- Probability they agree with each other = P(A)
- P(E) probability they would be expected to agree by chance
- 1 if judges agree. 0 if by chance. less than 0 disagree
- review example (example on slide 46)

Questions

1. How do we set significance level
2. Do example of sign test and (no kap)pa