# Standard Error & Confidence Interval

# Standard Error

- A particular kind of standard deviation
- Standard Error := standard deviation of the sampling distribution of a statistic
- Statistic := a function of a dataset (e.g., mean, median, variance, correlations, accuracy, f-score, ROUGE, BLEU)

- There is a nice closed form for computing standard error for sample mean (via Central Limit Theorem), but for most other statistics (e.g., median, variances, correlations, accuracy, f-score, ROUGE, BLEU), no general closed form formula available

# Bootstrap Estimate of Standard Error

- proposed by Efron (1979)
- an instance of "plug-in principle": plug-in sample statistics for unknown parameter values
- **Bootstrap Samples:** Using the empirical distribution (i.e., distribution of the dataset), *randomly generate* a number of new samples (a number of new datasets), where each sample (dataset) is of the same size as the original dataset.

# Bootstrap Estimate of Standard Error

- **Bootstrap Samples:** Using the empirical distribution (i.e., distribution of the dataset), *randomly generate* a number of new samples (a number of new datasets), where each sample (dataset) is of the same size as the original dataset.
- Compute the standard error of your statistic from these bootstrap samples. Recall **sample standard deviation** is defined by

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

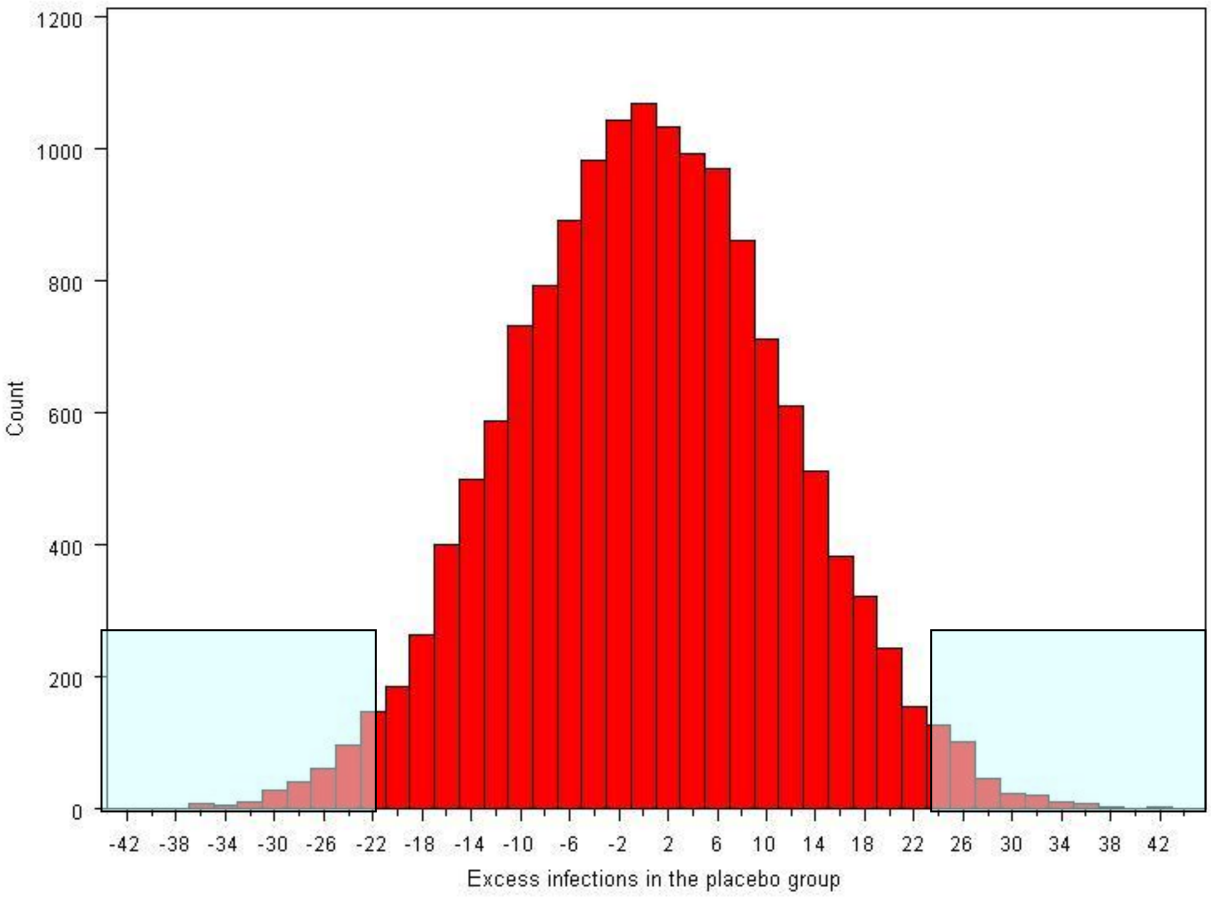
- Don't forget to use *N* – 1 instead of *N*! This correction is known as Bessel's correction.

# Confidence Interval

- Given confidence level (confidence co-efficient) $0 <= a <= 1$, we want to compute confidence interval [l, u] of a parameter x (a quantity we want to estimate) such that

  $$p(l < x < u) >= 1 – a$$

# Confidence Interval



# Confidence Interval

- Given confidence level (confidence co-efficient) $0 <= a <= 1$, we want to compute confidence interval [l, u] of a parameter x (a quantity we want to estimate) such that

$$p(l < x < u) >= 1 - a$$

- **Bootstrap Percentile Interval:**
  1. Generate bootstrap samples
  2. Sort the statistics computed from bootstrap samples
  3. Find the a/2 and 1-a/2 quantiles

# Hypothesis Testing

# Null Hypothesis / Alternative Hypothesis

- You have a baseline A and your own invention B
- B performs better than A by 1 % based on 10-fold cross validation
- How good is it?

- $H_o$ **Null Hypothesis:** A and B have the same performance.
  - that is, 1% difference is only a fluke
  - Skeptic’s point of view
- $H_a$ **Alternative Hypothesis:** B is indeed better than A

# Statistical Test

- A number of choices:
  - **Paired Student t-test**
  - **Sign test**
  - **Wilcoxon test**
  - **McNemar test**
  - **Permutation test**
  - **Bootstrap test**
- They all try to answer the following question:
  - should we ***reject*** Null Hypothesis **($H_o$)** or not?

# Statistical Test

- They all try to answer the following question:
  - should we ***reject*** Null Hypothesis **($H_o$)** or not?

  - whether we should accept null hypothesis?
  - whether we accept alternative hypothesis?
  - which hypothesis is better?

# Statistical Test

- They all try to answer the following question:
  - should we ***reject*** Null Hypothesis **($H_o$)** or not?

  - ~~whether we should accept null hypothesis?~~
  - ~~whether we accept alternative hypothesis?~~
  - ~~which hypothesis is better?~~

- Not rejecting Null Hypothesis… is the same as accepting Null Hypothesis?

# Statistical Test

- They all try to answer the following question:
  - should we ***<u>reject</u>*** Null Hypothesis **($H_o$)** or not?

  - ~~whether we should accept null hypothesis?~~
  - ~~whether we accept alternative hypothesis?~~
  - ~~which hypothesis is better?~~

- Not rejecting Null Hypothesis… is the same as accepting Null Hypothesis?

  ➔ NO! (it just means neither accepting nor rejecting)

# P-value

- They all try to answer the following question:
  - should we ***reject*** Null Hypothesis **($H_o$)** or not?
- We reject Null based on a threshold called **p-value**
- p-value: conditional probability of seeing MORE extreme results that what have been observed, conditional on the assumption that Null Hypothesis is true.
- typical p-value threshold is **0.05 (5%)**
- very small p-value == observation unlikely if Null is true

# Type I & II Error

- Type I Error:
  - When a test rejects a true null hypothesis
  - aka, False Positive
- Type II Error:
  - When a test fails to reject a false null hypothesis
  - aka, False Negative

- p-value bounds **Type I error**
- p-value: conditional probability of seeing MORE extreme results that what have been observed, conditional on the assumption that Null Hypothesis is true.

# Type I & II Error

- Type I Error:
  - When a test rejects a true null hypothesis
  - aka, False Positive
- Type II Error:
  - When a test fails to reject a false null hypothesis
  - aka, False Negative

- p-value bounds **Type I error**

➔ With typical p-value = 0.05 (5%), 1 out of 20 papers claims a scientific advance that is not there!

# Paired Student t-test

- Assumption: $D_i$ are independent and normally distributed
- $D_i$ is the difference between statistics of two different studies. For instance, the difference of accuracy (or f-score) of baseline and the proposed approach.
- Typically, we obtain N number of differences from N-fold cross validation.
- "paired" test in that the difference is computed from paired numbers that belong to the same evaluation setting (e.g., same fold in the N-fold cross validation)
- Null hypothesis := $\mu_D = 0$

# Paired Student t-test

$$t_D = \frac{\sqrt{N} m_D}{s_D}$$

- D is the set of differences of statistics (e.g., N difference in accuracies between 2 approaches with N-fold cross validation)
- $m_D$ is the sample mean of D
- $s_D$ is the sample standard deviation of D (with N-1 instead of N!)
- Above $t_D$ score follows t-distribution with N-1 degree of freedom, using which we can find the confidence interval efficiently.

# Paired Student t-test

$$t_D = \frac{\sqrt{N} m_D}{s_D}$$

- Above $t_D$ score follows t-distribution with N-1 degree of freedom (== $\nu$), using which we can find the confidence interval efficiently.

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

- Many tools available for which you only need to provide an array of paired numbers (R, various websites etc)

# Paired Student t-test: Issues to consider

- The **power** of a test is the probability of (correctly) rejecting the null hypothesis when it is in fact false.
- If D indeed satisfies the normality assumption, than T-test is very powerful in detecting statistical differences that other approaches may not able to detect.
- If D violates the normality assumption, or D is not independently distributed, or D has outliers or noises, then T-test is *not powerful* in detecting statistical differences. For those cases, consider non-parametric approaches instead.
- **Non-parametric approaches: sign-test, Wilcoxson test, NcNemar test, permutation test, bootstrap test**

# Parametric test

- Student t-test
- Paired Student t-test
- Wald test

➔ Assumes the data follows certain probabilistic distribution that are parameterized (e.g., normal distribution)

# Non-parametric test

- Sign test
- Wilcoxon signed-rank test
- NcNemar test
- permutation test
- bootstrap test

→ All of these assumes the data is ***independently*** distributed, but do not make assumptions based on well-known parametric distributions.

→ More ***powerful*** if the data do not follow certain parametric distributions (e.g., normal distribution)

# Sign Test & Wilcoxon test

- Let $V=v_1, \ldots, v_N$ and $U=u_1, \ldots u_N$ be the set of statistics of method A and method B respectively
  - E.g., they are prediction accuracy from N-fold cross validation.
- Let $D=d_1, \ldots, d_N$ be the difference between these ***paired*** statistics so that $d_i = v_i - u_i$

➔Student t-test & Wald test: whether the mean of $d_i$ is 0

➔Sign test: whether the number of cases where $d_i > 0$ is different from the number of cases where $d_i < 0$

➔Wilcoxon test: whether the median of the difference $d_i$ is 0.

This means, Sign test and Wilcoxon test depend only on the sign of the differences, not the magnitude!

# Sign Test

- Let $D=d_1, \ldots, d_N$ be the difference between these ***paired*** statistics so that $d_i = v_i - u_i$
- The null hypothesis H_0 of Sign Test := the sign of each $d_i$ is drawn from a bernoulli distribution so that
  - $p(d_i > 0) = 0.5$
  - $p(d_i < 0) = 0.5$
  - Cases such that $d_i = 0$ are ignored in this test
- Then pdf of k = the number of cases where $d_i > 0$ is

$$P(K=k) = \binom{M}{k} p^k (1-p)^{M-k}$$

  - where M is the number of non-zero cases in D, and p = 0.5
  - can compute p-value using cdf of binomial distribution

# McNemar Test

- Let $V=v_1, \ldots, v_N$ and $U=u_1, \ldots u_N$ be the set of statistics of method A and method B respectively.
- McNemar test is applicable when v_i and u_i are binary values: 0 or 1
- need to compute the "contingency table":

| | $v_i = 0$ | $v_i = 1$ | marginal |
|---|---|---|---|
| $u_i = 0$ | freq(0, 0) | freq(1, 0) | freq (*, 0) |
| $u_i = 1$ | freq(0, 1) | freq(1, 1) | freq(*, 1) |
| marginal | freq(0, *) | freq(1, *) | N |

# McNemar Test

| | $v_i = 0$ | $v_i = 1$ | marginal |
|---|---|---|---|
| $u_i = 0$ | freq(0, 0) | freq(1, 0) | freq (*, 0) |
| $u_i = 1$ | freq(0, 1) | freq(1, 1) | freq(*, 1) |
| marginal | freq(0, *) | freq(1, *) | N |

- The null hypothesis of McNemar test := marginal probabilities of each outcome (0 or 1) is the same over V and U. That is,
  - p(*, 0) = p(0, *)
  - p(1, *) = p(*, 1)

→ Intuitively, null hypothesis means freq(0, 1) and freq(1, 0) are close

→ Can map to binomial distribution with n = freq(0, 1) + freq (1, 0) and p=0.5

→ can also use chi-squared distribution, but not as exact as binomial if either freq(0, 1) or freq(1, 0) is small

# Bootstrap test

- Generate "bootstrap samples"
- Compute the confidence interval from the sorted list of statistics
- Reject the null hypothesis if the measured statistic is outside this confidence interval

# Bootstrap samples

Original Dataset
x_1, x_2, x_3, x_4, x_5

- Generate N bootstrap samples, where each bootstrap sample is the same size as the original dataset
- Each bootstrap sample contains data points that are **randomly sampled with replacement** from the original dataset

Bootstrap Sample 1
x_1, x_1, x_3, x_4, x_5

Bootstrap Sample 2
x_1, x_2, x_3, x_4, x_5

Bootstrap Sample 3
x_1, x_3, x_3, x_4, x_5

Bootstrap Sample 4
x_1, x_2, x_3, x_4, x_5

Bootstrap Sample 5
x_1, x_1, x_3, x_5, x_5

Bootstrap Sample 6
x_2, x_2, x_3, x_3, x_3

Bootstrap Sample 7
x_1, x_1, x_3, x_4, x_5

# Bootstrap samples

Original Dataset
x_1, x_2, x_3, x_4, x_5

- Compute N different statistics $V = v_1, \ldots, v_N$ using these N samples
- Compute the confidence interval (e.g., 95%) from the sorted list of V
- If the (assumed) statistic of null hypothesis is outside this confidence interval, reject the null hypothesis

Bootstrap Sample 1
x_1, x_1, x_3, x_4, x_5

Bootstrap Sample 2
x_1, x_2, x_3, x_4, x_5

Bootstrap Sample 3
x_1, x_3, x_3, x_4, x_5

Bootstrap Sample 4
x_1, x_2, x_3, x_4, x_5

Bootstrap Sample 5
x_1, x_1, x_3, x_5, x_5

Bootstrap Sample 6
x_2, x_2, x_3, x_3, x_3

Bootstrap Sample 7
x_1, x_1, x_3, x_4, x_5

# permutation test

- Generate a number of new samples (similarly as bootstrapping)
- By randomly permuting the predicted labels between the two approaches (baseline V.S. the proposed approach) == permutation on prediction
- How many different permutations?
  - $2^N$
  - → too many to enumerate all. Therefore, sample a subset using binomial distribution with p=0.5 and n=N
  - → confidence interval is computed from the sorted list of statistics

# permutation test V.S. bootstrapping test:

- permutation test:
  - sampling without replacement
  - sampling operates on the statistics (e.g. prediction) directly

- bootstrapping test:
  - sampling with replacement
  - sampling operates on the dataset
    - statistics are computed later on the generated bootstrap samples

# Parametric test (Recap)

- Student t-test
- Paired Student t-test
- Wald test

→ Assumes the data follows certain probabilistic distribution that are parameterized (e.g., normal distribution)

# Non-parametric test (Recab)

- Sign test
- Wilcoxon signed-rank test
- NcNemar test
- permutation test
- bootstrap test

→All of these assumes the data is ***independently*** distributed, but do not make assumptions based on well-known parametric distributions.

→More ***powerful*** if the data do not follow certain parametric distributions (e.g., normal distribution)