Evaluating Predictors

Thanks to: T Dietterich

Evaluating Hypotheses

Given limited data . . .

- Estimating h's true error
  - Sample Error ≠ True Error
  - Confidence intervals
  - Cross-Validation
- Comparing h_1 to h_2
  - Paired-t tests
  - McNemar's Test
- Appendix
  - Binomial distribution

Problems Estimating Error

- **Bias**: Difference between value of estimator and true value

  bias $\equiv E[\underline{err_S(h)}] - err_D(h)$
- If S is training set (used to produce h), $err_S(h)$ is optimistically biased
- To get unbiased estimate,
  - choose h and S independently
  - NOT $h := L(S)$
- **Variance**: Even with unbiased estimator, $\underline{err_S(h)}$ may still vary from $err_D(h)$
  - $\underline{err_S(h)}$ may be different from $\underline{err_{S'}(h)}$
  - especially if $|S|$, $|S'|$ small

Example

- Hypothesis h misclassifies
  12 of 40 examples in S

$$\underline{err_S(h)} \quad = \quad 12/40 \quad = \quad 0.30$$

- What is $err_D(h)$?
  - true error, over entire population?

Estimators

- Experiment: Given h
  1. Draw sample S of size $|S| = n$ according to distribution D
  2. Measure $err_S(h)$
- $err_S(h)$ is a random variable
  - (ie, result of experiment)
- $err_S(h)$ is unbiased estimator for $err_D(h)$
  - $E[err_S(h)] - err_D(h) = 0$

- Given (one) observation $err_S(h)$, what can we conclude about $err_D(h)$?

Confidence Intervals (informal)

- If
  - S contains n examples,
    drawn independently of h and each other
  - n > 30
- Then, w/ ≈ 95% probability

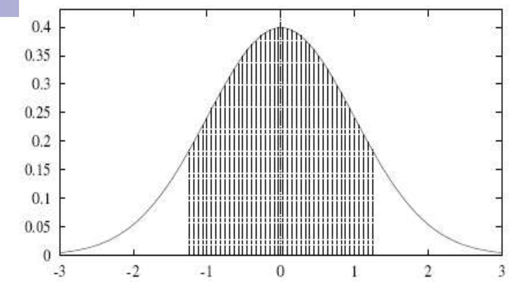
  $err_S(h)$ is in $err_{\mathcal{D}}(h) \pm 1.96 \sqrt{\frac{err_{\mathcal{D}}(h)\,(1 - err_{\mathcal{D}}(h))}{n}}$

- That is…

$$err_{\mathcal{D}}(h) \in \widehat{err}_S(h) \pm 1.96 \sqrt{\frac{err_{\mathcal{D}}(h)\,(1 - err_{\mathcal{D}}(h))}{n}}$$

$$\approx \widehat{err}_S(h) \pm 1.96 \sqrt{\frac{\widehat{err}_S(h)\,(1 - \widehat{err}_S(h))}{n}}$$

Elaboration



- If S contains n > 30 examples drawn independently of h, each other,
- Then can assume $\underline{err_S(h)} \sim N(\, err_D(h), \sigma^2)$

  $\underline{err_S(h)}$ drawn from Gaussian w/

  mean $\mu = err_D(h)$, var $\sigma^2 = err_D(h)(1 - err_D(h)) / n$

$\Rightarrow$ w/prob $\approx \alpha\%$,

$$\widehat{err}_S(h) \in [err_{\mathcal{D}}(h) - z_\alpha \cdot \sigma,\ err_{\mathcal{D}}(h) + z_\alpha \cdot \sigma]$$

ie, $\quad |\widehat{err}_S(h) - err_{\mathcal{D}}(h)| \leq z_N \cdot \sigma$

As $err_{\mathcal{D}}(h) \approx \widehat{err}_S(h)$, $\quad \sigma \approx \hat{s} = \sqrt{\dfrac{\widehat{err}_S(h)\ (1 - \widehat{err}_S(h))}{n}}$

$\Rightarrow$ w/prob $\approx \alpha\%$,

$$err_{\mathcal{D}}(h) \in [\widehat{err}_S(h) - z_\alpha \cdot \hat{s},\ \widehat{err}_S(h) + z_\alpha \cdot \hat{s}]$$

Example, con't

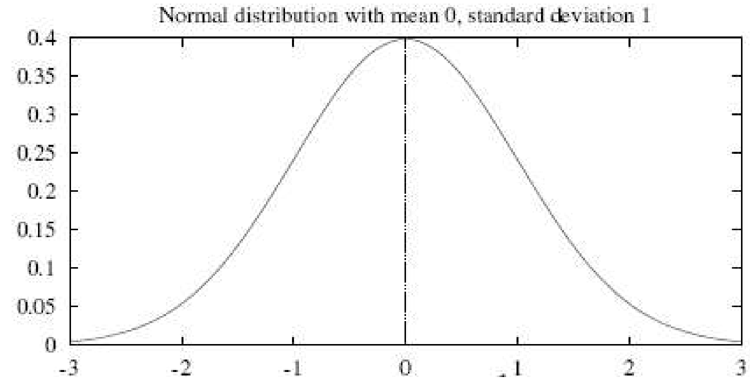
- For 12-of-40:
  - $\underline{err_S(h)} = 0.3$
  - $\hat{s} = \sqrt{(0.3 \times 0.7/ 40)} \approx 0.072$
- 95% confident that
  true error $err_D(h) \in \underline{err_S(h)} \pm 1.96\, \hat{s}$

$\Rightarrow err_D(h) \in [0.3 - 0.14, 0.3+0.14]$

- "Two-sided interval"
  - What about "one-sided interval"
    . . . likelihood that $err_D(h) < K$?

Normal Probability Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



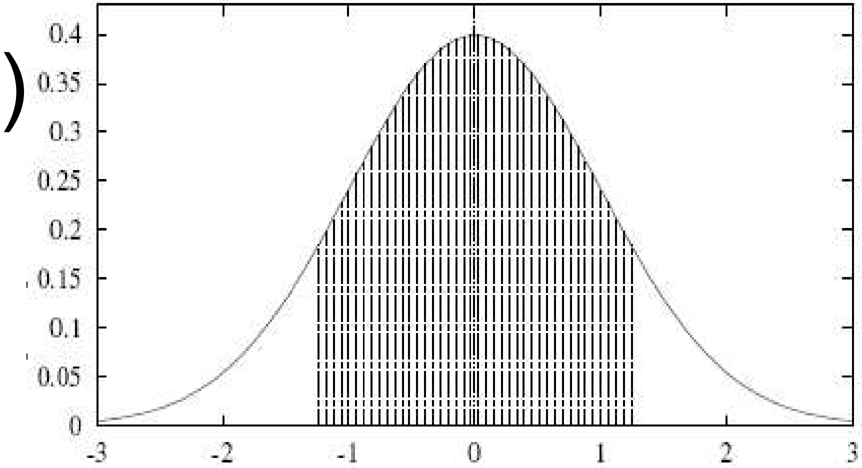
- $P(\, a \leq X \leq b\,)$
  $\equiv$ probability that X in interval (a, b) $= \int_a^b p(x)dx$
- $E[X] = \mu = \int_{-\infty}^{+\infty} x\; p(x)\, dx$
- $Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\, dx$
- $\sigma_X = \sqrt{Var(X)}$

Normal Probability Distribution

- 80% of area (probability) lies in $\mu \pm 1.28\sigma$

  $\in [\mu - 1.28\sigma, \mu + 1.28\sigma]$

- N% of area (probability) lies in $\mu \pm z_N \sigma$

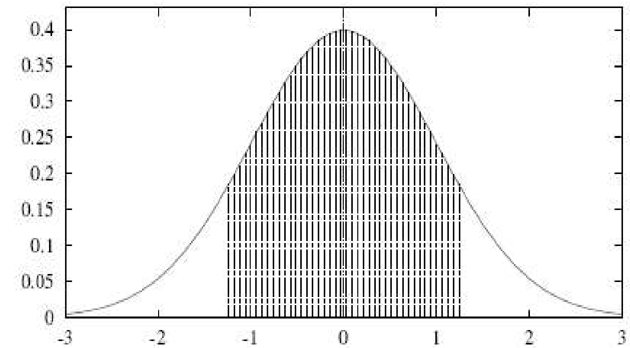


$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- If σ is small: Most of mass near mean μ

  If σ is large: Most of mass far from mean μ

One- vs Two- Sided Bounds



So far: "Constrain" μ to interval $[\hat{X} - z_n\sigma,\ \hat{X} + z_n\sigma]$

Eg, 80% confidence

$$err_{\mathcal{D}}(h) \ \in\ [\widehat{err}_S(h) - 1.28\hat{s},\ \widehat{err}_S(h) + 1.28\hat{s}]$$

- What is prob that $err_{\mathcal{D}}(h) \ \geq\ A$?

  Distribution is symmetric:

  ... 10% chance that

  $$err_{\mathcal{D}}(h) \ \in\ (-\infty,\ \widehat{err}_S(h) - 1.28\hat{s}]$$

  ... 10% chance that

  $$err_{\mathcal{D}}(h) \ \in\ [\widehat{err}_S(h) - 1.28\hat{s},\ +\infty)$$

  $\Rightarrow$ 90% chance

  $$err_{\mathcal{D}}(h) \ \in\ (-\infty,\ \widehat{err}_S(h) + 1.28\hat{s}]$$

One-Sided Bounds

If $\quad 100(1-\alpha)\%$ confident that $\mu \in [A, B]$,

Then $\quad 100(1-\frac{\alpha}{2})\%$ confident that $\mu \in [A, +\infty)$
ie, $\quad \mu \geq A$

and $\quad 100(1-\frac{\alpha}{2})\%$ confident that $\mu \in (-\infty, B]$
ie, $\quad \mu \leq B$

- Confidence of one-sided error
  is TWICE the confidence of two-sided!

Eg, For 12-of-40:
- 95% confident $err_D(h) \in [0.3 - 0.14, 0.3+0.14]$
- 97.5% confident $err_D(h) \leq 0.3+0.14$

Central Limit Theorem

- Let $Y_1, \ldots Y_n$ be set of iid r.v.s
  (independent, identically distributed random variables)
  all drawn from same arbitrary distribution
  with mean μ and finite variance σ^2.
  - sample mean $\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- **Central Limit Theorem**
  As $n \to \infty$, $\hat{Y} \sim N(\mu, \sigma^2/n)$

$$\boxed{\frac{\hat{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)}$$

- Distribution governing $\hat{Y}$ approaches Normal distribution, w/ mean μ, variance σ^2/n
  - Y_i from ANY distribution, just same $\forall\ Y_i$
  - Typically apply when $n > 30$

Calculating Condence Intervals General Procedure

- 1. Identify parameter p to estimate
  - $err_D(h)$
- 2. Choose an estimator
  - $err_S(h)$
- 3. Determine prob distr of estimator
  - $err_S(h)$ ~ Binomial distribution,
  - … approximated by Normal when n > 30
- 4. Find interval (L, U) such that N% of probability mass falls in the interval
  - Use table of z_N values

Truth. . .

- $\widehat{err}_S(h) = \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$

  where $Y_i = \begin{cases} 1 & \text{if } i^{th} \text{ instance mislabeled} \\ 0 & \text{otherwise} \end{cases}$

- $\widehat{err}_S(h)$ is ASYMPTOTICALLY normal

  As $|S| \to \infty$, $\quad \widehat{err}_S(h) \sim \mathcal{N}(err_{\mathcal{D}}(h), \sigma^2)$

  $$\sqrt{|S|}\frac{\widehat{err}_S(h) - err_{\mathcal{D}}(h)}{\sigma} \sim \mathcal{N}(0,1)$$

  assuming σ^2 is known!

- If σ^2 not known, then
  - $\hat{\sigma} := \sqrt{\frac{\widehat{err}_S(h)\ (1-\widehat{err}_S(h))}{|S|-1}}$
  - $\sqrt{|S|}\ \frac{\widehat{err}_S(h) - err_{\mathcal{D}}(h)}{\hat{\sigma}} \sim t_{|S|-1}$
  - "students t" distribution

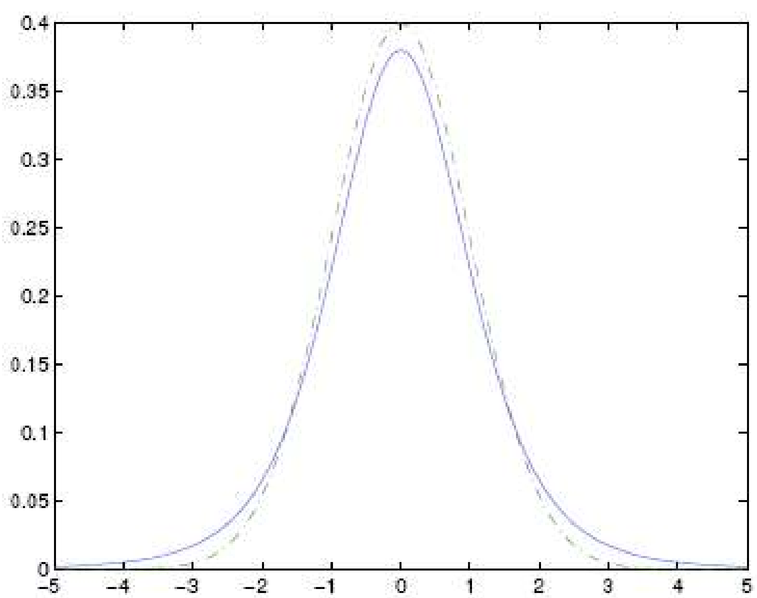
Students t Distribution

- t distribution like unit normal N(0, 1) but larger spread (longer tail)

$\Rightarrow$ interval (for given α) is larger

... additional uncertainty due to unknown variance

$\lim_{k\to\infty} t_{\alpha,k} = z_\alpha$



	Confidence Level N			
	90%	95%	98%	99%
$\nu = 2$	2.92	4.30	6.96	9.92
$\nu = 5$	2.02	2.57	3.36	4.03
$\nu = 10$	1.81	2.23	2.76	3.17
$\nu = 20$	1.72	2.09	2.53	2.84
$\nu = 30$	1.70	2.04	2.46	2.75
$\nu = 120$	1.66	1.98	2.36	2.62
$\nu = \infty$	1.64	1.96	2.33	2.58
z_N	1.64	1.96	2.33	2.58

IIa. Difference Between Hypotheses

Test h_1 on sample S_1, test h_2 on S_2

1. Pick parameter to estimate
   - $d = err_D(h_1) - err_D(h_2)$
2. Choose an estimator
   - $\underline{d} = \underline{err_S(h_1)} - \underline{err_S(h_2)}$

   (Btw, $E[\underline{d}] = d$)
3. Determine prob distr of estimator

$$\sigma_{\hat{d}} \approx \sqrt{\frac{\widehat{err}_{S_1}(h_1)\,(1-\widehat{err}_{S_1}(h_1))}{|S_1|} + \frac{\widehat{err}_{S_2}(h_2)\,(1-\widehat{err}_{S_2}(h_2))}{|S_2|}}$$

(Diff of 2 Normals is Normal)

4. Find interval (L, U) s.t. N% of probability mass in interval

$$\hat{d} \pm z_N \sqrt{\frac{\widehat{err}_{S_1}(h_1)\,(1-\widehat{err}_{S_1}(h_1))}{|S_1|} + \frac{\widehat{err}_{S_2}(h_2)\,(1-\widehat{err}_{S_2}(h_2))}{|S_2|}}$$

(Tighter bound [better] if use $S_1 = S_2$)

Example (con't)

- Spse $\underline{err}_A(h_A) = 0.3$; $\underline{err}_B(h_B) = 0.4$; given $|S_A| = 100 = |S_B|$
- As $\underline{d} = \underline{err}_A(h_A) - \underline{err}_B(h_B) = 0.1 > 0$

  h_B appears better that h_A

- **Q:** Is h_B truly better than h_A . . .

  ie, Is $err_D(h_B) < err_D(h_A)$?

  … ie what is prob that $d < 0$
  given observed $\underline{d} = 0.1$?

- **A:** Assume null-hypothesis: $d = \mu_d < 0$.
  - What is chance that $P(d = 0.1 \mid \underline{d} < 0)$?
    . . . bounded by chance that estimate $\underline{d}$ is OFF by > 0.1
  - . . . $\underline{d}$ in 1-sided interval $\underline{d} \in [\ \mu_d + 0.1, \infty)$

Examples . . . Hypothesis Testing

- What is chance that $\underline{d} \in [\ \mu_d + 0.1, \infty)$

- Here: $\underline{\sigma}_d \approx 0.061$.
  - With prob > 0.95, $\underline{d} < \underline{d} + 1.64\ \underline{\sigma}_d$
- $\Rightarrow$ Given $\underline{d} = 0.1$,

  95% confident that prob that $d > 0$

  … ie, $err_D(h_A) > err_D(h_B)$
- **Hypothesis Test:**
  - Accept hyp $err_D(h_A) \leq err_D(h_B)$ with confidence 0.95
  - Reject null hyp (that $err_D(h_A) > err_D(h_B)$)

    at 1 – 0.95 = 0.05 level of significance

Paired-t Test to compare h_A, h_B

Given: data T; alg's h_A; h_B; confidence α:

- 1. Partition data into *k* disjoint test sets $\{ T_1, T_2, \ldots, T_k \}$
  of ≈equal size (size ≥ 30)
- 2. For i = 1 .. k, do $\delta_i := err_{Ti}(h_A) - err_{Ti}(h_B)$
- 3. Le

  $$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

  (empirical estimate of standard deviation)
- 4. Return α% confidence estimate for d: $\underline{\delta} \pm t_{\alpha,k-1}\, s_{\underline{\delta}}$

- Hypothesis test:
  Is $\underline{\delta} + t_{\alpha,k-1}\, s_{\underline{\delta}} > 0$?
- Note: When each δ_i is ≈ Normally distributed… $\underline{\delta}$ ~ “Students T”

IIb. Comparing Two Classifiers

- **Goal**: decide which of two classifiers h_1 vs h_2 has lower error rate
- **Method**: Run both on same test data set, recording following numbers:

		classified by h_A correct	classified by h_A incorrect
classified by h_B	correct	n_{00}	n_{10}
	incorrect	n_{01}	n_{11}

McNemar's Test

		classified by h_A	
		correct	incorrect
classified by h_B	correct	n_{00}	n_{10}
	incorrect	n_{01}	n_{11}

$$M = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} > \chi^2_{1,\alpha}$$

- *M* is distributed approximately as χ^2 w/ 1 degree of freedom
- For 95% confidence: $\chi^2_{1,\ 0:95} = 3.84$
- So if $M > 3.84$

  reject null hyp that

  "h_A, h_B have same error rate"

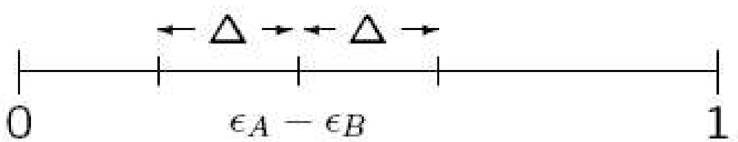
Confidence Interval... Difference Between Two Classifiers

- $p_{ij} = \frac{n_{ij}}{n}$ be 2x2 contingency table, as probabilities

$$
\begin{aligned}
SE &= \sqrt{\frac{p_{01} + p_{10} + (p_{01} - p_{10})^2}{n}} \\
p_A &= p_{10} + p_{11} \\
p_B &= p_{01} + p_{11} \\
\Delta &= 1.96(SE + \frac{1}{2n})
\end{aligned}
$$

- 95% confidence interval on difference in true error $\epsilon_A - \epsilon_B$ between two classifiers:

$$(p_A - p_B) \in [\epsilon_A - \epsilon_B - \Delta,\ \epsilon_A - \epsilon_B + \Delta]$$



Estimate Diff Between Two Alg's: the 5x2CV F test

for i from 1 .. 5 do
  %perform 2-fold cross-validation
  split S evenly and randomly into S_1, S_2
  for $j \in \{1,2\}$ do
    Train algorithm A on S_j, measure error rate $p_A^{(i,j)}$
    Train algorithm B on S_j, measure error rate $p_B^{(i,j)}$
    $p_i^{(j)} = p_A^{(i,j)} - p_B^{(i,j)}$ *% diff in err rates on fold* j

  $\bar{p}_i \ := \ \frac{p_i^{(1)} + p_i^{(2)}}{2}$ *% ave diff in err rates in iteration* i
  $s_i^2 = \left(p_i^{(1)} - \bar{p}_i\right)^2 + \left(p_i^{(2)} - \bar{p}_i\right)^2$ *% var in diff, for iter* i

$F \ := \ \frac{\sum_i \bar{p}_i^2}{2 \sum_i s_i^2}$

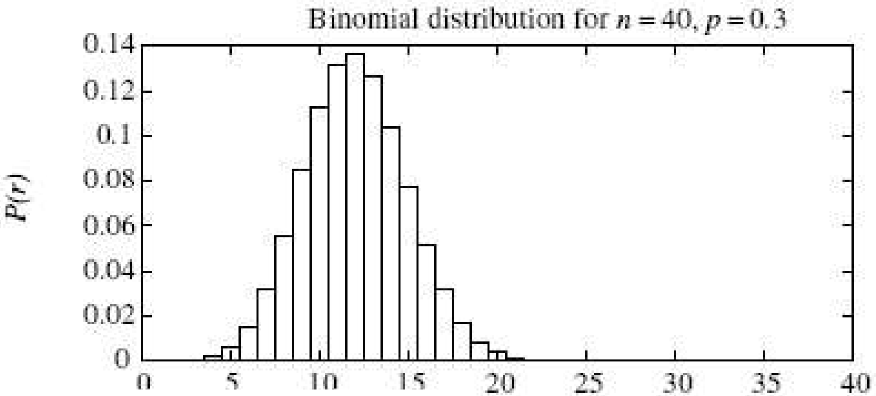
- If F > 4.47, then
  - with 95% confidence,
  - reject null hypothesis that
    alg's A and B have the same error rate
  when trained on data sets of size m/2

Other Topics

- Hypothesis testing, in general
- “False discovery rate”
  …permutation tests, . . .
- Prior knowledge of Distributions
- ROC curves
- ANOVA
- Running “experiments” to obtain data . . .
- . . .

$err_S(h)$ is a Random Variable

- Rerun experiment w/ different randomly drawn S (of size |S| = n)
- Prob of observing r misclassified examples:



$$P(r) = \binom{n}{r} err_{\mathcal{D}}(h)^r \, (1-err_{\mathcal{D}}(h))^{n-r}$$

$$\binom{n}{r} \equiv \frac{n!}{r!(n-r)!}$$

Binomial Probability Distribution

- If $p = P(heads)$, prob of r heads in n coin flips

Let: $Y_i = \begin{cases} 1 & i^{th} \text{ flip is heads} \\ 0 & \text{otherwise} \end{cases}$

$$X = \sum_{i=1}^{n} Y_i$$

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

- $E[X] \equiv$ Expected value of X:

$$\equiv \sum_{r=0}^{n} r \times P(X = r) = n \times p$$

- $Var(X) \equiv$ Variance of X

$$\begin{aligned} &\equiv E[(X - E[X])^2] \\ &= \sum_{r=0}^{n} (r - E[X])^2 \times P(X = r) \\ &= np(1-p) \end{aligned}$$

- $\sigma_X \equiv$ standard deviation of X

$$\equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}$$

Binomial Distribution, con't

- If $p = P(head)$, prob of r heads in n coin flips

Let: $Y_i = \begin{cases} 1 & i^{th} \text{ flip is head} \\ 0 & \text{otherwise} \end{cases}$

$$S = \sum_{i=1}^{n} Y_i \qquad \bar{Y} = \frac{S}{n}$$

- $E[\bar{Y}] \equiv$ Expected value of $\bar{Y}$:

$$= \frac{1}{n} E[S] = \frac{n \times p}{n} = p$$

- $Var(\bar{Y}) \equiv$ Variance of $\bar{Y}$

$$= E[(\tfrac{S}{n} - E[\tfrac{S}{n}])^2] = \tfrac{1}{n^2} E[(S - E[S])^2]$$

$$= \tfrac{1}{n^2} n\, p\,(1-p) = \tfrac{p\,(1-p)}{n}$$

- $\sigma_{\bar{Y}} \equiv$ standard deviation of $\bar{Y}$

$$\equiv \sqrt{Var(\bar{Y})} = \sqrt{\frac{p\,(1-p)}{n}}$$

Proofs

$$E[S] = \sum_{r=0}^{n} r \times P(r,n)$$

$$= \sum_{r=1}^{n} r \times \frac{n!}{r!(n-r)!} p^r \ (1-p)^{n-r}$$

$$= \sum_{r=1}^{n} \frac{n \times (n-1)!}{(r-1)!(n-r)!} p \times p^{r-1} \ (1-p)^{n-r}$$

$$= n\,p \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!((n-1)-(r-1))!} p^{r-1} \ (1-p)^{(n-1)-(r-1)}$$

$$= n\,p \sum_{s=0}^{n-1} \frac{(n-1)!}{s!((n-1)-s)!} p^s \ (1-p)^{(n-1)-s}$$

$$= n\,p\,(p + (1-p))^{n-1} = n\,p$$

$$Var(S) = E[(S-\mu)^2] = E[S^2 - 2\mu S + \mu^2]$$

$$= E[S^2] - 2\mu E[S] + \mu^2 = E[S^2] - E[S]^2$$

Binomial Approximates Normal Distribution

- $\widehat{err}_S(h)$ follows a *Binomial* distribution:
  - Mean $\mu_{\widehat{err}_S(h)} = err_{\mathcal{D}}(h)$
  - Standard deviation $\sigma_{\widehat{err}_S(h)}$

$$\sigma_{\widehat{err}_S(h)} = \sqrt{\frac{err_{\mathcal{D}}(h)\ (1-err_{\mathcal{D}}(h))}{n}}$$

- Can approximate as *Normal* distribution:
  - Mean $\mu_{\widehat{err}_S(h)} = err_{\mathcal{D}}(h)$
  - Standard deviation

$$\sigma_{\widehat{err}_S(h)} \approx \sqrt{\frac{\widehat{err}_S(h)\ (1-\widehat{err}_S(h))}{n}}$$