Stat 13 Lecture 22 comparing proportions

- Estimation of population proportion
- Confidence interval ; hypothesis testing
- Two independent samples
- One sample, competitive categories (negative covariance)
- One sample, non-competitive categories (usually, positive covariance)

An Example

Assume the sample is simple random.

1996 US	Pre-	election	polls			Elec	tion	res ult
	n	Clinton	Dole	Perot	other	Cl	Do	Pe
New Jersey	1000	51%	33%	8%	8%	53%	36%	9%
New York	1000	59%	25%	7%	9%	59%	31%	8%
Connecticut	1000	51%	29%	11%	9%	52%	35%	10

Does the poll result significantly show the majority favor Clinton in New Jersey? For Dole, is there is a significant difference between NY and Conn ? Find a 95% confidence interval for the difference of support between Clinton and Dole in New Jersey?

Do you play

- Tennis ? Yes, No
- Golf? Yes, No
- Basketball? Yes, No

	Yes	No
T	30%	70%
G	25%	75%
B	40%	60%

n=100 persons are involved in the survey

Gene Ontology
200 genes randomly selected

cytoplasm	nucleus	others	unknown
60	35	20	100

Central limit theorem implies that binomial is approximately normal when n is large

- Sample proportion is approximately normal
- The variance of sample proportion is equal to p(1-p)/n
- If two random variables,X, Y are independent, then variance of (X-Y) = var (X) + var(Y)
- If two random variables, X,Y are dependent, then variance of (X-Y)=var (X) + var(Y)-2cov(X,Y)
- May apply the z-score formula to obtain confidence interval as done before.

One sample, Competitive categories

- X=votes for Clinton, Y=votes for Dole
- Suppose sample size is n=1, then only three possibilities $P(X=1, Y=0)=p_1$; $P(X=0,Y=1)=p_2$; $P(X=0,Y=0)=1-p_1-p_2$
- $E(X)=p_1$; $E(Y)=p_2$
- $Cov(X,Y) = E(X-p_1)(Y-p_2) = (1-p_1)(0-p_2)p_1 + (0-p_1)(1-p_2)p_2 + (0-p_1)\ (0-p_2)\ (1-p_1-p_2)$
- $= -p_1p_2$, **which is negative**
- In general, $cov(X,Y)= \mathbf{-n\ p_1\ p_2}$; **therefore**
- **Var (X/n - Y/n)=**n^{-2} (Var X + Var Y + $2np_1p_2$)
- $=n^{-2}(np_1(1-p_1) + np_2(1-p_2) + 2np_1p_2)=$
- $(p_1 + p_2 - p_1^2 - p_2^2 + 2p_1p_2)/n = \mathbf{(p_1+p_2-(p_1-p_2)^2)/n}$

Formula for confidence interval

- Let $\hat{p}_1 = X/n$, $\hat{p}_2 = Y/n$
- Then the interval runs from
- $\hat{p}_1 - \hat{p}_2 - z\ sd(p_1 - p_2)$, to
- $\hat{p}_1 - \hat{p}_2 + z\ sd(p_1 - p_2)$
- Where sd is the square root of variance, plug in the variance formula