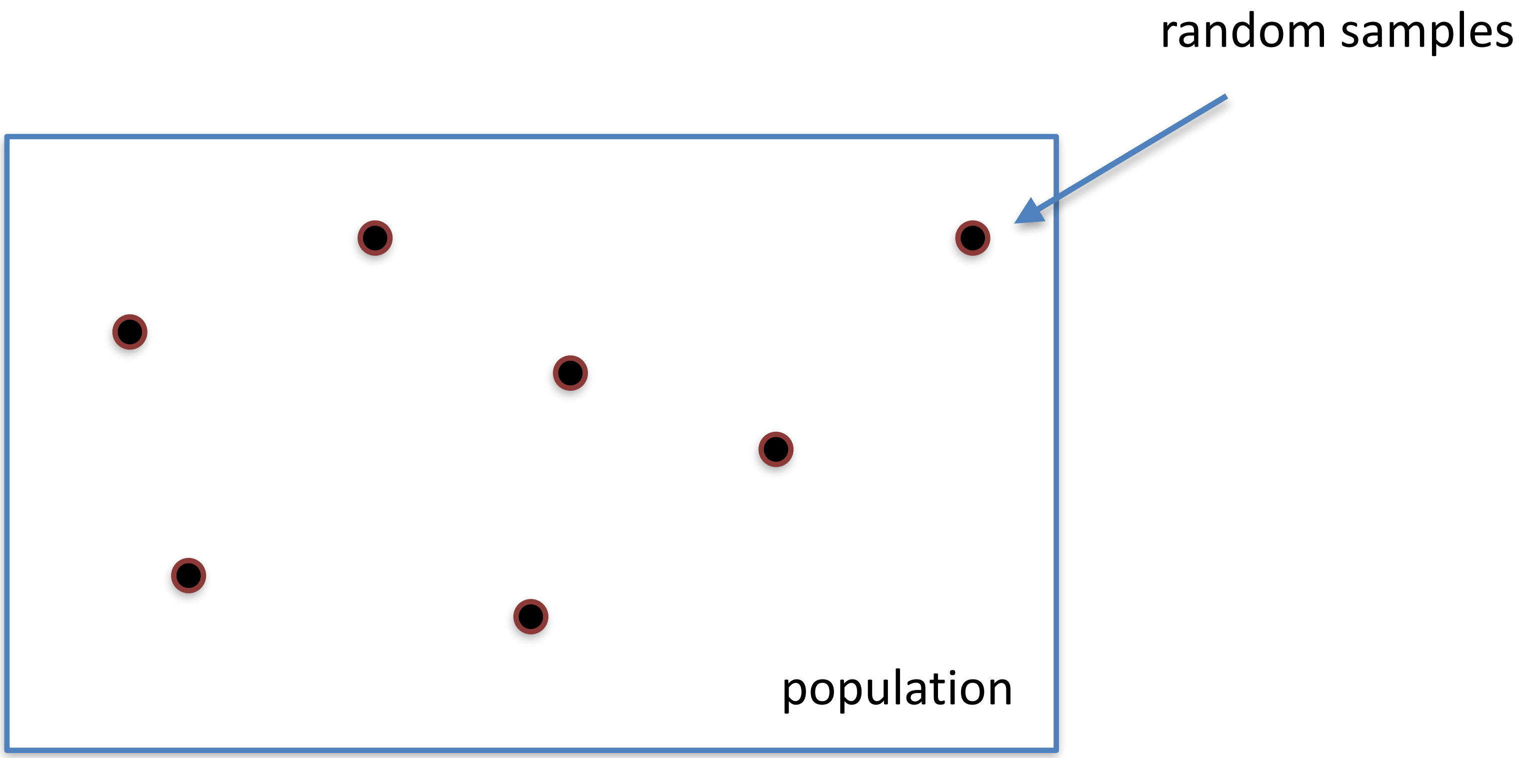
Final Review

Fall 2013

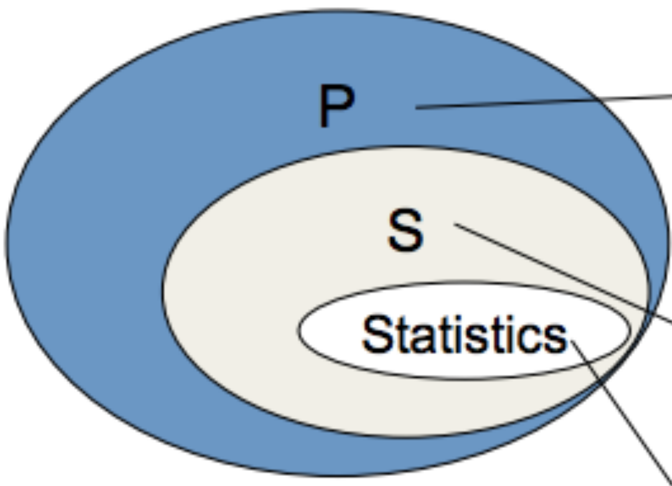
Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech

Random sampling model



random samples: $x_1, \ldots, x_n$

For example, we use digital thermometer to measure body temperature for 5 times, we obtain a sequence.
If we do this experiment the next day, we get a different sequence of measures.
The result of the measurement is a sequence of **random samples (also called data)**.



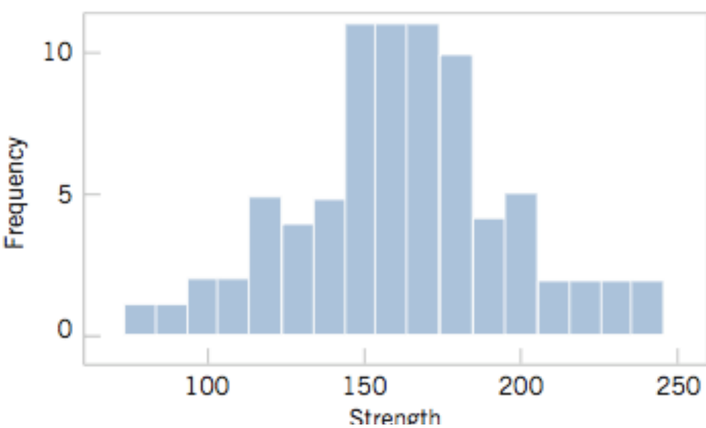
Population : all possible observations

Sample : observed measurements taken from population

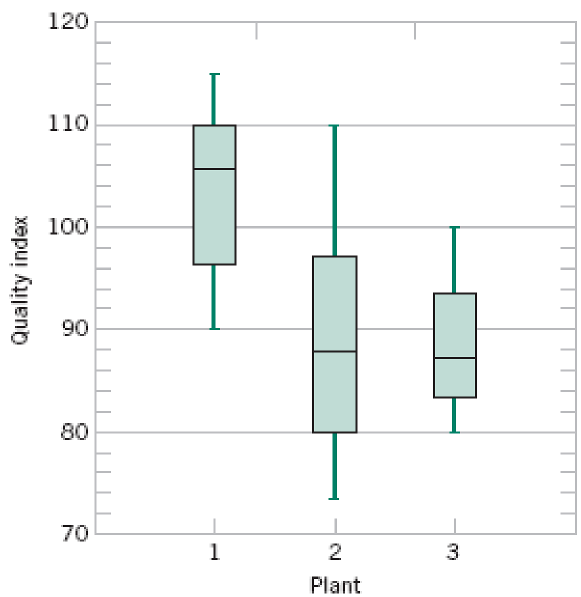
Statistics : useful summary of the sample data

Descriptive statistics

- Quantitative values
    - provides simple summaries about samples
- plot



Histogram

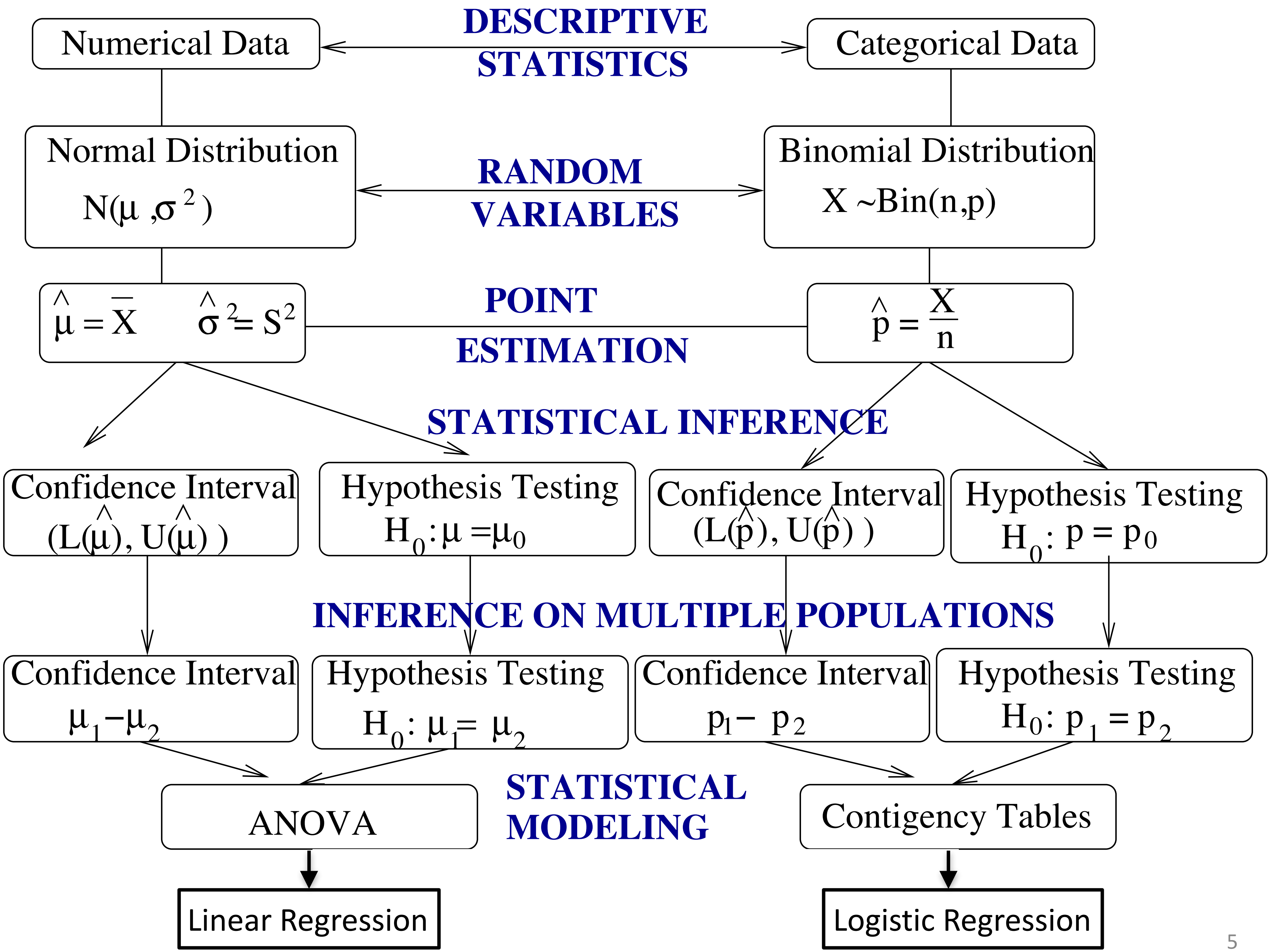


Box plot

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

Stem & Leaf diagram



Numerical Data
DESCRIPTIVE STATISTICS
Categorical Data
Normal Distribution
$N(\mu, \sigma^2)$
RANDOM VARIABLES
Binomial Distribution
$X \sim Bin(n,p)$
$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2$
POINT ESTIMATION
$\hat{p} = \frac{X}{n}$
STATISTICAL INFERENCE
Confidence Interval
$(L(\hat{\mu}), U(\hat{\mu}))$
Hypothesis Testing
$H_0: \mu = \mu_0$
Confidence Interval
$(L(\hat{p}), U(\hat{p}))$
Hypothesis Testing
$H_0: p = p_0$
INFERENCE ON MULTIPLE POPULATIONS
Confidence Interval
$\mu_1 - \mu_2$
Hypothesis Testing
$H_0: \mu_1 = \mu_2$
Confidence Interval
$p_1 - p_2$
Hypothesis Testing
$H_0: p_1 = p_2$
ANOVA
STATISTICAL MODELING
Contigency Tables
Linear Regression
Logistic Regression

Data summary

- Samples $x_1, x_2, \cdots, x_n$
- Sample mean $\bar{x} = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n$
- Sample median
  - 1) rank samples from smallest to largest

    $y_1, y_2, \cdots, y_n$

  - 2) odd number of samples, median = $y_{(n+1)/2}$

    even number of samples, median =

    $(y_{(n-1)/2} + y_{(n-1)/2})/2$

- Sample range = largest - smallest
- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

- Sample quartile x_p: pth quartile is such that p-percent of samples are smaller than x_p

- upper quartile
- lower quartile

- Inter quartile range (IQR) = upper quartile - lower quartile

Sampling distribution

- Distribution of the **statistics** we come up (above)
- Sampling distribution extremely useful for determining
  - forms of confidence interval
  - hypothesis test

Sampling distribution: summary

	Sample mean	*Sample variance*
Form	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
sample i.i.d. normal Known variance	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$	$\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$
Unknown variance	*large n, approximately normal as above*	*large n, approximately normal*

Other common sampling distribution

Sample proportion $\hat{p} = \frac{X}{n}$	Standardized sample mean, known variance $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$	Standardized sample mean, unknown variance $\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$
Exact: $n\hat{p} \sim BIN(n,p)$	**Exact** $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N\left(0, \quad 1 \right)$	**Exact** $\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}$
Large sample: $\hat{p} \sim N(np, np(1-p))$		

Two sample

Difference in sample mean, known variance	*Difference in sample mean, unknown (but identical) variance,*	*Proportion of sample variance*
$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ $\sim N(0,\ 1)$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $\sim t_{n_1+n_2-1}$	$\dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$

$$S_p^{\ 2} = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2 + \sum_{i=1}^{n_1}\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}$$

Statistical methods

- Point estimator
- Confidence interval
- Hypothesis test

- Two sample test (two populations)
- ANOVA (more than two populations)

- Linear regression

Point estimator

- Mean of estimator: unbiased
- Variance of estimator
- Mean Square Error (MSE)
  - MSE = biase^2 + variance

- Method of finding point estimators
  - method of moment
  - maximum likelihood

Confidence interval

- Point estimator: a single value for estimated parameter
- Confidence interval: an interval such that true parameter lies in

- [a, b] contains true parameter with probability $1-\alpha$
- then [a, b] is the $1-\alpha$ confidence interval

Typical forms of k

- k = upper cutting point * variance of point estimator

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

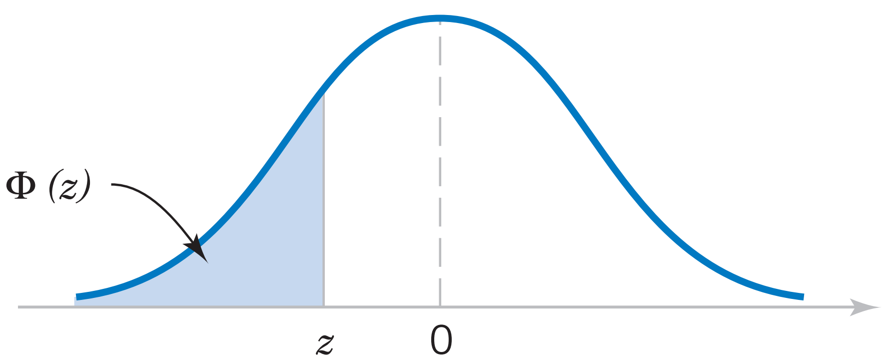
$$\left(\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2},n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2},n-1}\right)$$

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\left(1-\hat{p}\right)/n}, \quad \hat{p} - z_{\alpha/2}\sqrt{\hat{p}\left(1-\hat{p}\right)/n}\right)$$

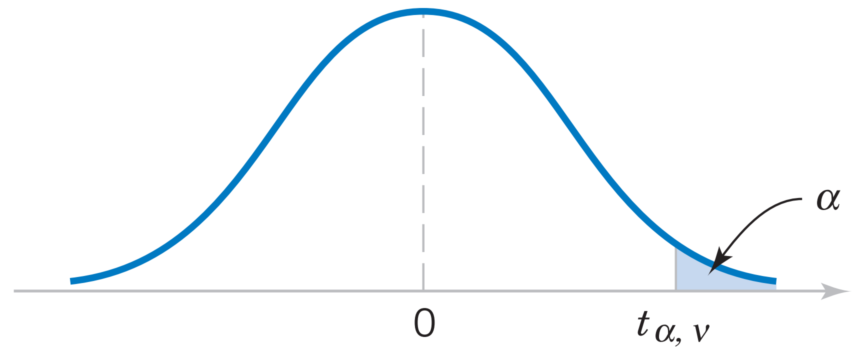
- width of confidence interval determined by sample size and confidence level

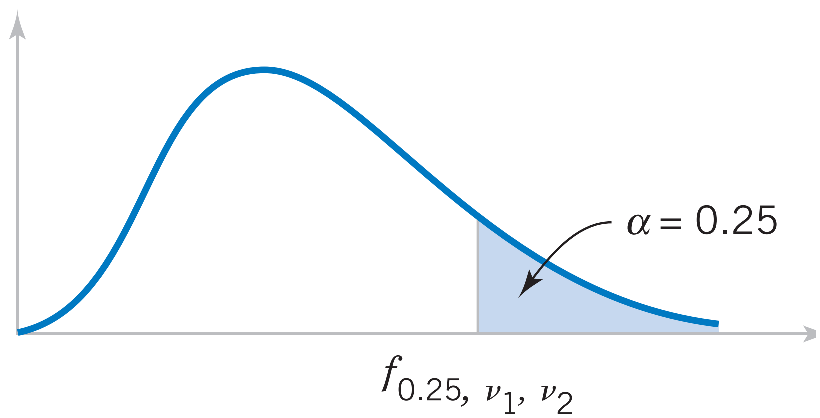
Tails etc

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$



CDF





Upper cutting point
(also called "percentage point" in textbook)

Forms of confidence intervals

- Two-sided interval

[point estimator - k, point estimator + k]

- One-sided interval

[point estimator + k, infinity]

or

[-infinity, point estimator - k]

- k specifies width of confidence interval

Hypothesis test

- Use data to test two contradicting statements
  - H_0: null hypothesis
  - H_1: alternative hypothesis
- Two approaches
  - Fixed confidence level
    - Form: reject H_0 when **test statistic** falls out of **thresholds**
  - p-value
    - probability of observing something more “extreme” than data

Procedure of hypothesis test (sec. 9.1.6)

1. Set the significance level (.01, .05, .1)
2. Set null and alternative hypothesis
3. Determine other parameters
4. Decide type of the test
   - test for mean with known variance (z-test)
   - test for mean with unknown variance (t-test)
   - test for sample proportion parameter
6. Use data available:
   - perform test to reach a decision
   - and report p-value

Summary: test for mean

Null Hypothesis

$$H_0 : \mu = \mu_0$$

Test Statistic

$$\overline{x}$$

Significance level: α

Alternative Hypothesis	Known Variance **H0 is rejected if**	Unknown Variance **H0 is rejected if**
$H_1 : \mu \neq \mu_0$	$\lvert \overline{x} - \mu_0 \rvert > z_{\alpha/2} \sigma / \sqrt{n}$	$\lvert \overline{x} - \mu_0 \rvert > t_{\alpha/2, n-1} s / \sqrt{n}$
$H_1 : \mu > \mu_0$	$\overline{x} > \mu_0 + z_\alpha \sigma / \sqrt{n}$	$\overline{x} > \mu_0 + t_{\alpha, n-1} s / \sqrt{n}$
$H_1 : \mu < \mu_0$	$\overline{x} < \mu_0 - z_\alpha \sigma / \sqrt{n}$	$\overline{x} < \mu_0 - t_{\alpha, n-1} s / \sqrt{n}$

Test for sample proportion

Null Hypothesis

$$H_0 : p = p_0$$

Test Statistic

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Significance level: α

Alternative Hypothesis	**H0 is rejected if**
$H_1 : p \neq p_0$	$\left\| \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \right\| > z_{\alpha/2}$

Two sample test: mean

For the following hypothesis test

$$H_0 : \mu_1 - \mu_2 = \Delta$$
$$H_1 : \mu_1 - \mu_2 \neq \Delta$$

Reject H_0 when

$$\left| \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \right| > t_{\alpha/2}$$

Two-sample test: sample proportion

For two-sided test, $H_0 : p_1 = p_2$

$$H_1 : p_1 \neq p_2$$

reject H_0 when

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \right| > z_{\alpha/2}$$

Analysis of variance

- Multiple populations
- Analyze difference in their means

Table 13-2 Typical Data for a Single-Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\dots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\dots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\dots$	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

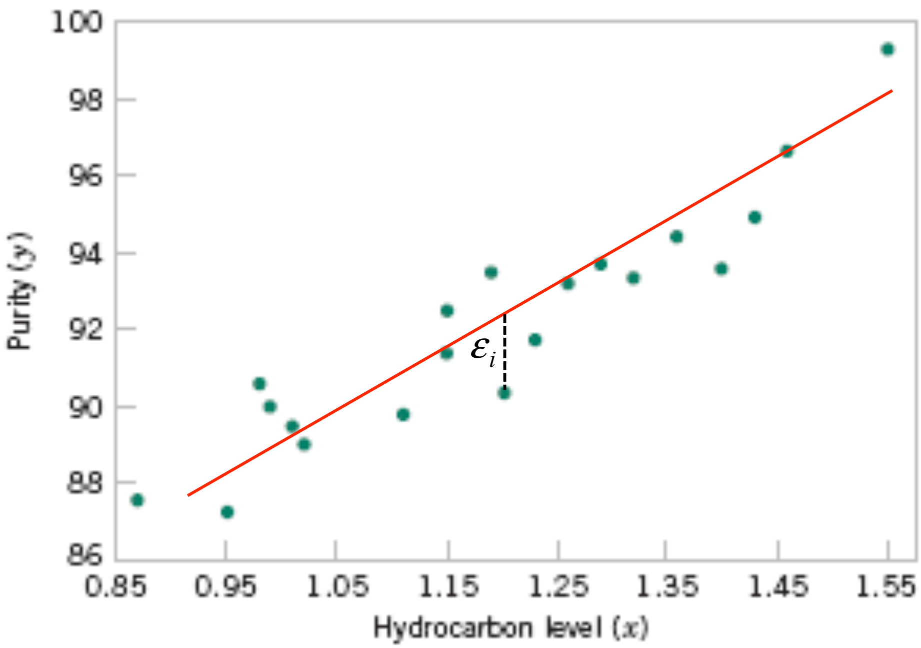
$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/[a(n-1)]} = \frac{MS_{\text{Treatments}}}{MS_E} \tag{13-7}$$

We would reject H_0 if

$$F_0 > F_{\alpha,a-1,a(n-1)}$$

Linear regression

- Simple linear regression



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

Intercept

Slope

Random error

Fitted coefficients

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (11\text{-}10)$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad (11\text{-}11)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

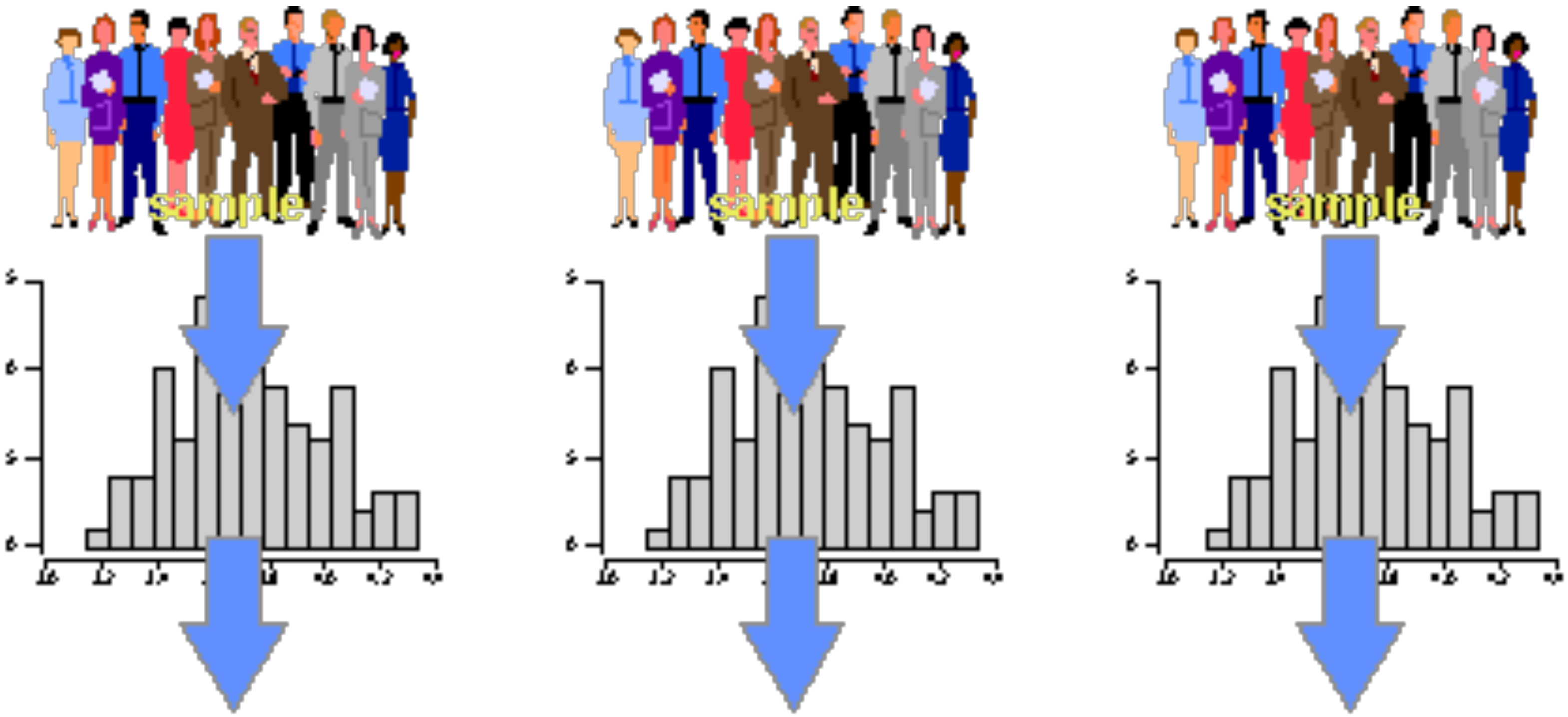
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted (estimated) regression model

Model diagnosis

- Plot residuals
- Use R and read the output
- For simple and multiple linear regression: we are going to rely on R to do the calculations

Finally…

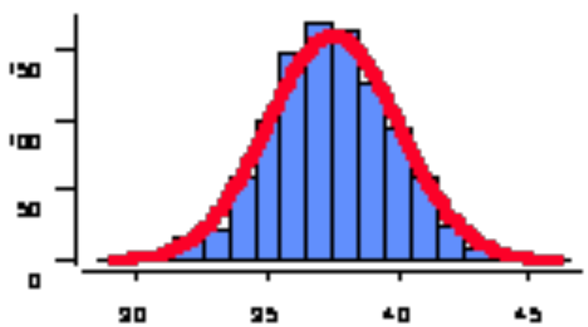


Average

Average

Average

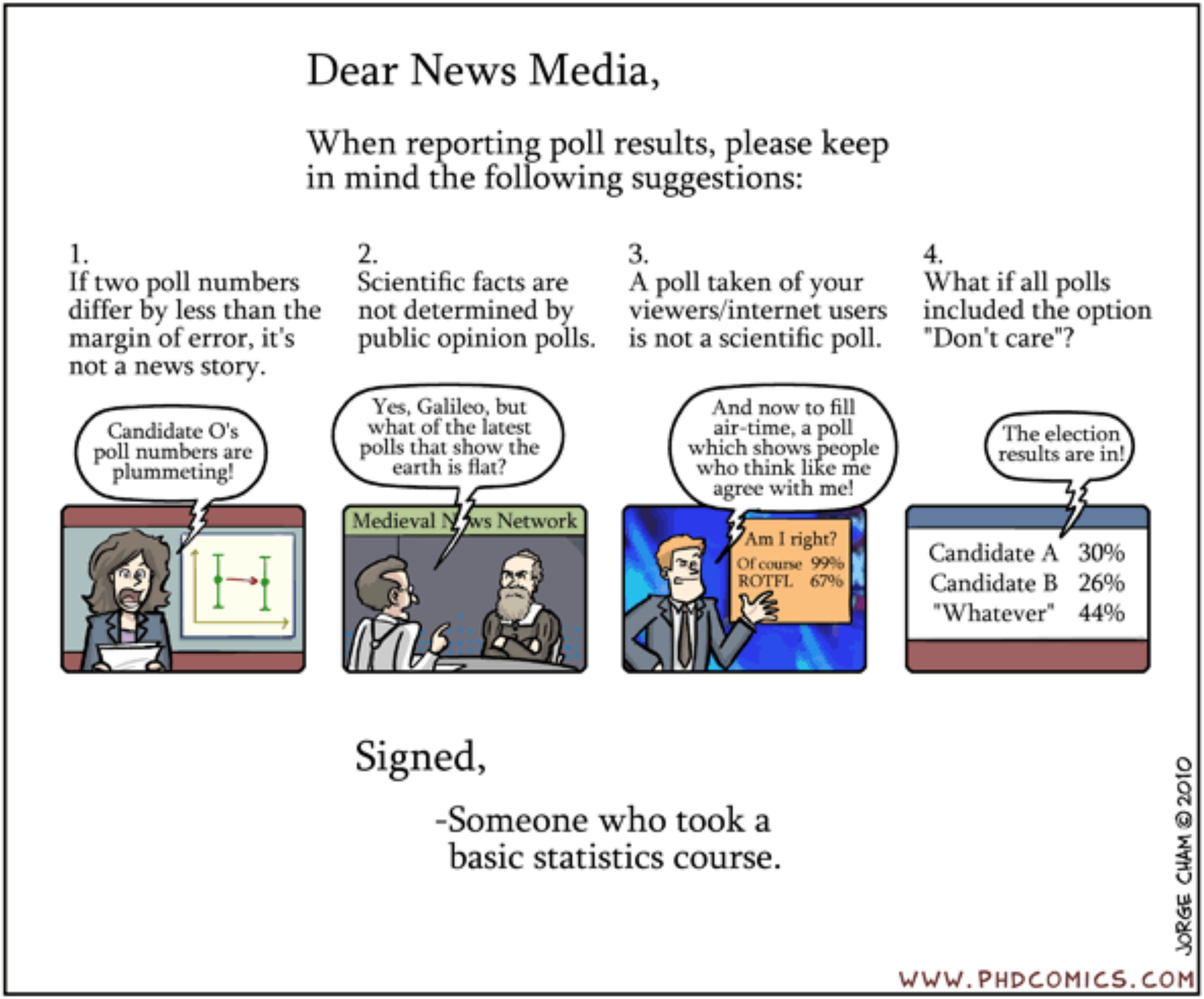
The Sampling Distribution…



…is the distribution of a statistic across an infinite number of samples

Finally…

- What statistics is about?
- Fit model using data (e.g. distributions)
- Use model to make inferences
  - estimation
  - hypothesis testing
  - prediction (e.g. using linear regression)
- Why model is useful?
  - report findings from data
  - systematically quantify **uncertainty**



Dear News Media,
When reporting poll results, please keep in mind the following suggestions:
1. If two poll numbers differ by less than the margin of error, it's not a news story.
Candidate O's poll numbers are plummeting!
2. Scientific facts are not determined by public opinion polls.
Yes, Galileo, but what of the latest polls that show the earth is flat?
Medieval News Network
3. A poll taken of your viewers/internet users is not a scientific poll.
And now to fill air-time, a poll which shows people who think like me agree with me!
Am I right?
Of course 99%
ROTFL 67%
4. What if all polls included the option "Don't care"?
The election results are in!
Candidate A 30%
Candidate B 26%
"Whatever" 44%
Signed,
-Someone who took a basic statistics course.
JORGE CHAM © 2010
WWW.PHDCOMICS.COM