Purpose and Nature of Sampling

- **Nature:** only incomplete view of a true picture available in real life
- **Purpose:** to describe a clearly defined population on the basis of sample information
- **Statistics:** Various functions of the data may be used to calculate measures, each of which is a reflection of some special feature of the population. These sample measure are called **statistics**

1 ©Argon Chen

Measure of Central Tendency (Location)

- **Sample Mean:** of a set of numbers **(lower case in expressions)** $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Note:
  - sample mean is a "statistic" and is not a **true mean but an estimate of mean**

2

Measure of Dispersion

- **Sample Variance** of the set $x_1, x_2, \ldots, x_n$ of **numerical observations**, denoted by s^2 is given by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\boxed{n-1}}$$

Degree of freedom=n-1
Why not n?

- Sample variance is a statistic used to estimate variance
- The **sample standard deviation**, denoted by s, is the positive square root of the sample variance

3

Degree of Freedom (DoF)

- DoF: the number of independent pieces of information
- DoF=the number of values free to vary in calculation of a statistic
- Suppose there are n observations x_i, i=1,…, n.
  To calculate the sample mean, there are n independent pieces of information available for calculation of the statistic:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{\boxed{n}}$$

DoF=n

4

DoF of $x_i - \bar{x}$

- x_i–average is called ***residual*** or ***centered-measure*** and **Σ(x_i–average)=0**
- Let n=3; x_1=1, x_2=2, x_3=3; average=(1+2+3)/3=2
- That is, if we know x_1-average(=–1) and x_2–average(=0), then x_3-average must be known(= 1)

  → If you know two of x_i-average, you know the third. The number of values free to vary is 2!
- How many independent pieces of information do we have about (x_i-average)? Answer: 2
- For x_i–average, i=1,…,n, there are only n–1 pieces of independent information that are free to vary because the average has taken one piece of information and the nth value is subject to zero sum.

Taking average of (x_i–average)^2

- Again, let n=3 and x_1=1, x_2=2, x_3=3
- x_1–average=–1; x_2–average=0, and x_3–average=1
- What is the good estimate of E(x–mean)^2?

  Σ(x_i–average)^2/3 or Σ(x_i–average)^2/2?
- Σ(x_i–average)^2/3=2/3 or Σ(x_i–average)^2/2=1 more plausible?

  Σ(x_i–average)^2/2!

6

More on the DoF of $(x_i - \text{average})^2$

- Average is used to estimate the mean
- What if "median" is used to estimate of mean?
- Again, let n=3 and x_1=1, x_2=2, x_3=3:median=x_2
- DoF of $(x_i - \text{median})^2$?
- Since x_2 has been used to estimate the mean, only x_1–median=–1 and x_3–median=1 are left to estimate the distance from the center!

➢ The estimate of E(x-mean)^2? $\Sigma(x_i - \text{median})^2/2$

7

Sample Covariance

- **There must be correlations among measurements. For example, the higher blood pressure level often comes with the higher cholesterol level.**
- **The co-variation of two measurements is measured by sample covariance: (the trend that one is larger then the other is larger or one is smaller then the other is smaller)**

$$Cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i)}{n-1}$$

- $\boldsymbol{X}\uparrow$, $\boldsymbol{Y}\uparrow \Rightarrow$ **Cov>0: positively correlated**
- $\boldsymbol{X}\uparrow$, $\boldsymbol{Y}\downarrow \Rightarrow$ **Cov<0: negatively correlated**

8

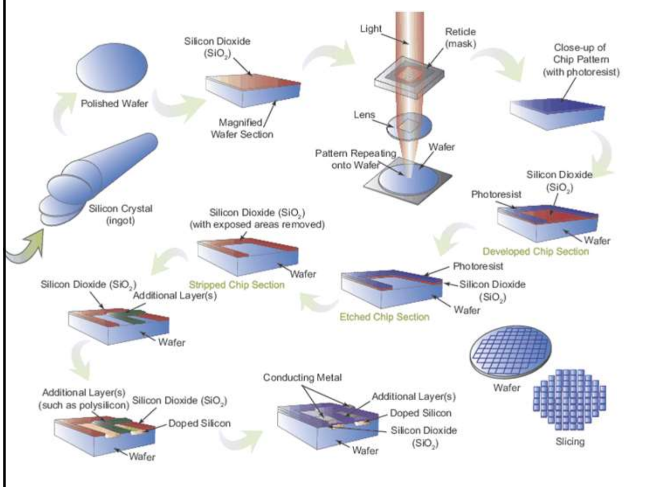
Sample Correlation Coefficient

- **Correlation coefficient is defined a σ (-1≤ρ_{xy}≤1):**

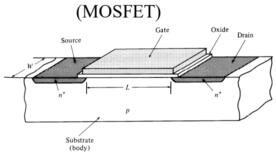
$$\rho_{\mathrm{xy}} = \frac{s_{\mathrm{xy}}}{s_{\mathrm{x}} s_{\mathrm{y}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

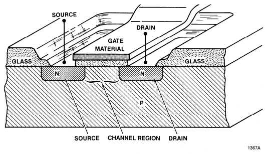
- **-1≤ ρ_{xy} <0: negatively correlated; 0< ρ_{xy} ≤1: positively correlated**
- **ρ_{xy} =0: no correlation; ρ_{xy} =±1: perfect correlation**
- **Let x'=(x–x-bar)/s_x and y'=(y–y-bar)/s_y**
  → $Cov(x', y')=\rho_{xy}$

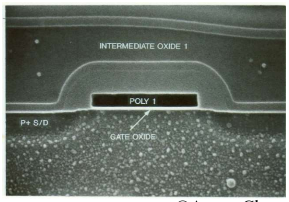
Semiconductor Manufacturing



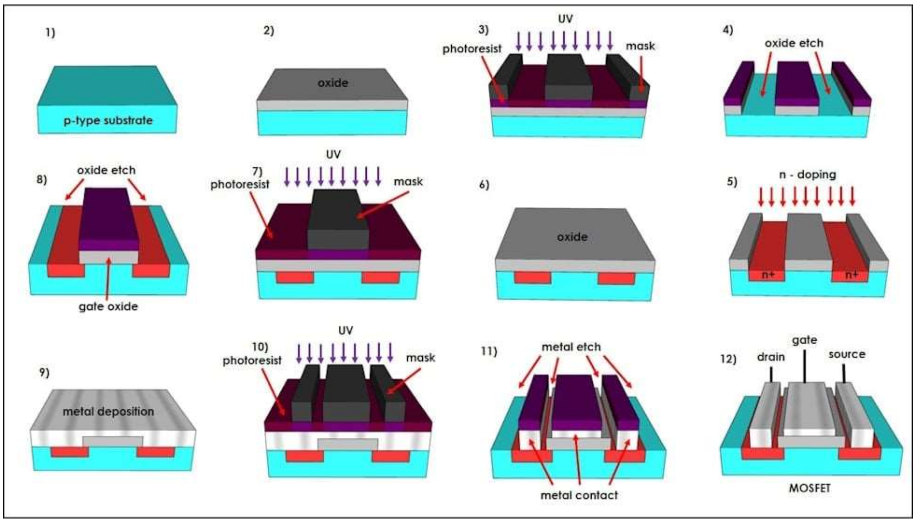
Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET)





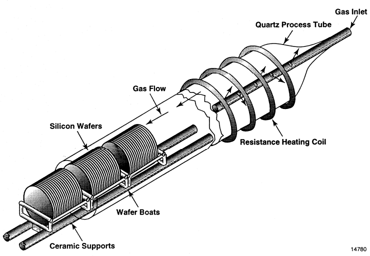


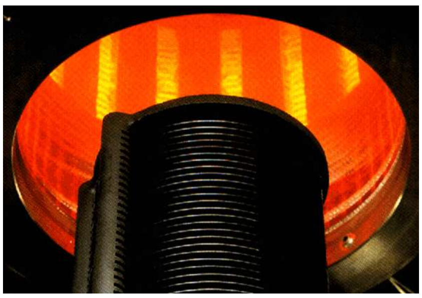
MOSFET Fabrication

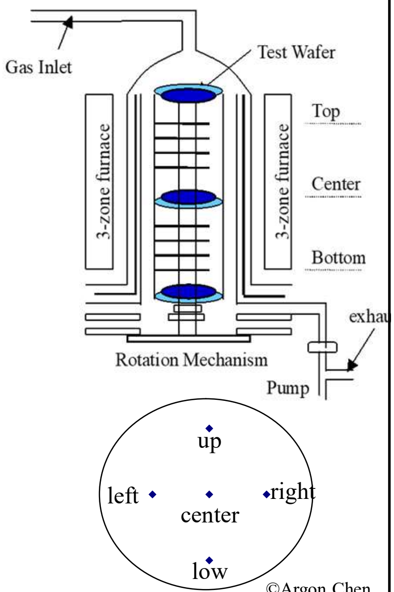


11 ©Argon Chen

Oxidation Furnace







12 ©Argon Chen

Thickness of SiO_2 Layer

- Example: 85 readings of SiO_2 average thickness (with target thickness=350Å) are collected

程 式	Thickness	Thickness (Top Zone)					Thickness (Center Zone)					Thickness (Bottom Zone)					
Recipe	target	Upper	Center	Lower	Left	Right	Upper	Center	Lower	Left	Right	Upper	Center	Lower	Left	Right	Average
73SC35	350	349	349	352	347	353	352	352	353	351	354	355	351	356	352	359	352.33
73SC35	350	347	349	352	349	353	354	353	354	352	354	353	351	359	352	354	352.40
73SC35	350	350	352	349	354	353	355	356	355	354	356	353	357	362	355	361	354.80
73SC35	350	346	348	351	346	351	352	352	352	349	353	351	351	357	352	357	351.20
73SC35	350	347	349	350	349	351	351	352	351	350	353	350	353	356	353	357	351.47
73SC35	350	345	348	350	346	350	350	352	350	348	352	351	351	353	351	356	350.20
73SC35	350	349	350	353	349	353	354	354	354	351	355	351	352	360	353	362	353.33
73SC35	350	348	349	351	347	351	352	352	352	350	353	352	350	357	350	357	351.40
73SC35	350	349	351	353	348	353	354	354	353	352	354	353	353	359	352	358	353.07
73SC35	350	351	353	355	354	358	352	353	353	351	354	353	353	355	351	360	353.73
73SC35	350	351	352	354	353	357	353	352	353	351	353	353	353	354	351	359	353.27
73SC35	350	347	347	349	346	349	351	352	352	351	353	349	350	359	352	356	350.87
73SC35	350	347	349	351	347	350	352	352	351	348	351	350	351	361	350	359	351.27
73SC35	350	347	349	351	346	351	352	352	350	353	352	349	351	360	351	358	351.47
73SC35	350	348	349	349	348	351	352	353	353	351	354	351	351	360	351	360	352.07
73SC35	350	347	349	350	346	350	351	351	350	352	351	349	350	359	351	357	350.87
73SC35	350	348	342	353	350	351	352	352	352	354	354	369	366	374	369	374	357.33
73SC35	350	356	355	357	354	356	338	339	338	342	340	356	351	361	357	361	350.73
73SC35	350	352	350	355	350	356	346	344	345	347	345	355	354	366	356	366	352.47
73SC35	350	348	346	349	345	353	344	343	343	343	345	353	352	363	353	362	349.47
73SC35	350	351	347	352	348	357	345	345	344	346	345	354	353	364	354	364	351.27
73SC35	350	351	349	353	349	356	346	346	345	346	347	354	354	360	356	360	351.47
73SC35	350	350	349	353	350	358	348	350	346	346	350	358	356	366	357	366	353.53
73SC35	350	350	349	352	350	358	348	350	346	347	350	358	355	365	357	366	353.40
73SC35	350	352	351	355	350	358	348	348	346	346	349	356	353	365	356	364	353.13
73SC35	350	350	350	350	349	353	344	344	345	343	346	333	346	338	335	338	344.27
73SC35	350	347	348	351	347	350	337	337	338	336	339	351	349	361	351	361	346.87
73SC35	350	349	350	353	349	355	344	344	345	344	346	350	349	348	349	350	348.33
73SC35	350	349	349	351	349	355	345	345	346	345	348	352	352	364	354	361	351.00
73SC35	350	349	350	354	350	354	344	345	345	345	346	351	357	363	353	361	351.13
73SC35	350	348	350	353	348	354	345	345	345	344	347	352	352	364	354	362	350.87
73SC35	350	352	350	352	348	355	349	349	348	348	350	349	351	363	352	362	351.87
73SC35	350	351	349	351	347	354	349	350	348	350	349	349	351	362	351	361	351.47
73SC35	350	347	350	351	347	353	348	348	348	345	349	347	348	361	349	358	349.93
73SC35	350	347	351	350	348	353	348	349	348	345	348	347	348	360	349	351	349.47
73SC35	350	349	350	351	348	349	354	350	360	350	359	352	351	354	349	350	351.73
73SC35	350	344	344	348	343	349	346	344	344	342	345	347	343	353	344	353	345.93
73SC35	350	353	353	358	352	359	350	351	351	352	354	354	354	362	355	364	354.80
73SC35	350	351	351	354	351	354	342	342	343	340	344	352	350	360	350	361	349.67
73SC35	350	352	352	354	351	354	349	349	349	347	351	350	350	352	358	352	351.33
73SC35	350	350	349	353	349	354	349	348	348	347	350	351	350	360	352	360	351.33
73SC35	350	352	352	355	351	355	349	349	347	350	349	331	344	336	334	336	346.00
73SC35	350	349	351	354	350	355	348	348	347	347	349	348	349	359	349	358	350.73
73SC35	350	350	350	354	349	355	349	349	349	346	350	353	351	362	352	362	352.07
73SC35	350	348	348	352	347	351	339	340	341	338	342	351	349	358	351	358	347.53
73SC35	350	350	351	353	349	355	350	350	348	347	351	351	352	361	352	360	352.00
73SC35	350	349	349	352	348	354	349	350	349	349	351	351	353	360	353	361	351.87
73SC35	350	351	351	353	351	354	343	343	343	342	345	355	355	262	252	364	337.60
73SC35	350	352	353	355	352	355	342	344	345	342	345	356	355	364	356	364	352.00
73SC35	350	351	352	355	350	357	350	350	349	347	350	351	350	359	351	359	352.07
73SC35	350	350	350	353	350	357	350	350	349	348	350	352	351	360	351	360	352.07
73SC35	350	353	352	357	352	358	355	354	355	353	356	355	352	362	354	362	355.33
73SC35	350	350	351	353	349	353	349	349	350	347	350	349	347	357	348	358	350.67
73SC35	350	350	350	354	350	355	351	351	350	348	352	350	349	359	350	359	351.87
73SC35	350	348	348	355	348	353	350	350	350	348	351	349	348	357	349	358	350.80
73SC35	350	348	348	345	348	349	345	346	355	345	356	348	347	348	349	351	348.53
73SC35	350	348	348	346	347	349	345	346	354	346	355	348	347	349	350	351	[illegible]
73SC35	350	346	346	349	344	351	348	345	348	346	348	345	346	355	346	355	347.87
73SC35	350	346	346	349	345	351	348	348	348	346	349	343	346	356	346	356	348.20

13 ©Argon Chen

Frequency Distribution

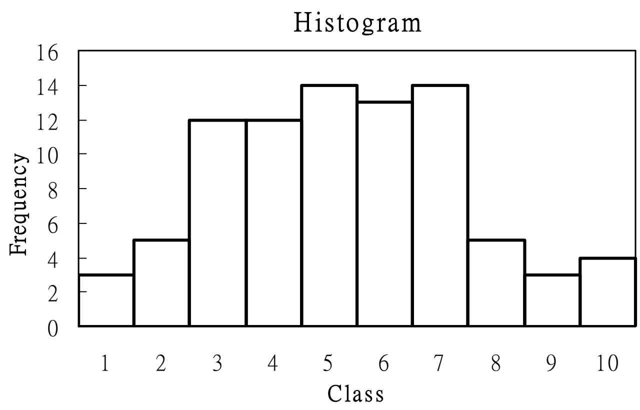
- Example: frequency distribution is calculated for 85 readings of SiO_2 average

	Class Interval	Frequency=*fi*	Relative Freq.= *fi* /Tota
1	~346.5	3	0.0353
2	346.6~347.5	5	0.0588
3	347.6~348.5	12	0.1412
4	348.6~349.5	12	0.1412
5	349.6~350.5	14	0.1647
6	350.6~351.5	13	0.1529
7	351.6~352.5	14	0.1647
8	352.6~353.5	5	0.0588
9	353.6~354.5	3	0.0353
10	354.6~	4	0.0471
	Total	85	

14 ©Argon Chen

Histograms

- Example: 350Å SiO_2 thickness data



15 ©Argon Chen

The Box Plot

The Five-Number Summary:

Min ----- Q_1----- Median ----- Q_3 ----- Max

•Divides the data into 4 sets containing an equal number of measurements.

•A quick summary of the data distribution.

•Use to form a **box plot** to describe the **shape** of the distribution and to detect **outliers.**

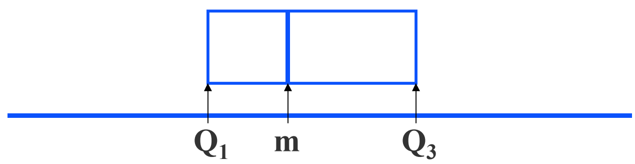
16 ©Argon Chen

Constructing a Box Plot

✓Calculate Q_1, the median, Q_3 and IQR($=Q_3-Q_1$).

✓Draw a horizontal line to represent the scale of measurement.

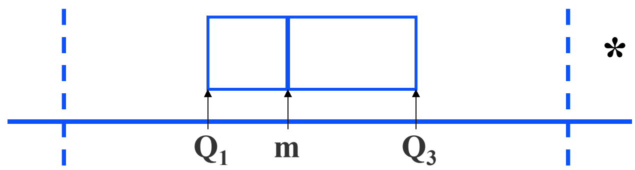
✓Draw a box using Q_1, the median m, Q_3.



Constructing a Box Plot

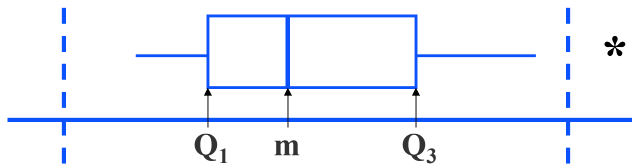
✓Isolate outliers by calculating
- ✓Lower fence: $Q_1-1.5$ IQR (or $Q_1-3(m-Q_1)$)
- ✓Upper fence: $Q_3+1.5$ IQR (or $Q_3+3(Q_3-m)$)

✓Measurements beyond the upper or lower fence are outliers and are marked with *.



Constructing a Box Plot

✓Draw "whiskers" connecting the largest and smallest measurements that are NOT outliers to the box.



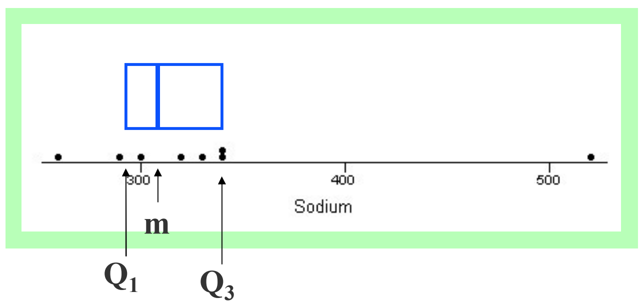
19 ©Argon Chen

Example

Amt of sodium in 8 brands of cheese:

260 290 300 320 330 340 340 520

$Q_1 = 292.5$ $m = 325$ $Q_3 = 340$



20 ©Argon Chen

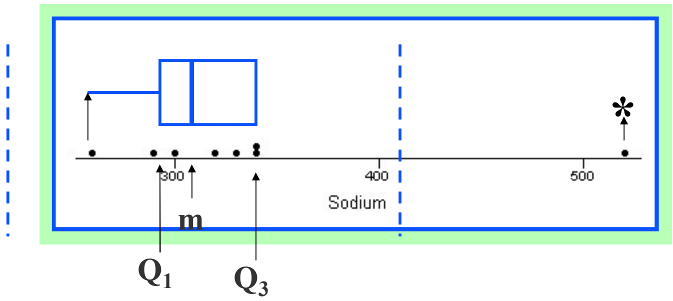
Example

IQR = 340-292.5 = 47.5

Lower fence = 292.5-1.5(47.5) = 221.25

Upper fence = 340 + 1.5(47.5) = 411.25

Outlier: x = 520



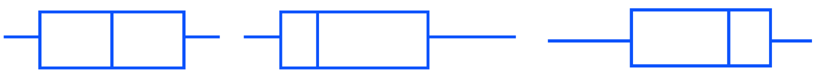
21 ©Argon Chen

Interpreting Box Plots

✓Median line in center of box and whiskers of equal length—symmetric distribution

✓Median line left of center and long right whisker—skewed right

✓Median line right of center and long left whisker—skewed left



22 ©Argon Chen

Statistic

- A **statistic** is any function of the random variables constituting one or more samples, provided that the function does not depend on any unknown parameter values

  Examples: sample mean, sample variance
- Sample data:
  - A **sample** = A set of sample observations $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n]$ and **sample size**=n
  - A sample **observation** = A piece of data vector $\boldsymbol{x}_i=[x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}]$

What does Statistics do?

- **Point estimate:**
  - To estimate the parameters of the probability models with sample data
  - To evaluate how good the estimators are
- **Hypothesis test:**
  - To check/test whether the model parameter(s) has changed.
  - To evaluate how good the tests are (two types errors?)
- **Mathematical modeling of sampling statistics for performance evaluation:**
  - Modeling the "point estimate" and "hypothesis testing" for their performance evaluation

Point Estimate

- A **point estimate** of a parameter θ is a single number that can be regarded as the most plausible value of θ. A point estimate is obtained by selecting **a suitable statistic** and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ

A Point Estimate of Mean: Sample Mean (Average)

- A statistic of point estimate, say sample mean, is a random variable. For example

1st sampling: $\bar{x}_{1st} = \dfrac{\sum_{i=1}^{n} x_{1st,i}}{n}$ 2nd: $\bar{x}_{2nd} = \dfrac{\sum_{j=1}^{n} x_{2nd,j}}{n}$

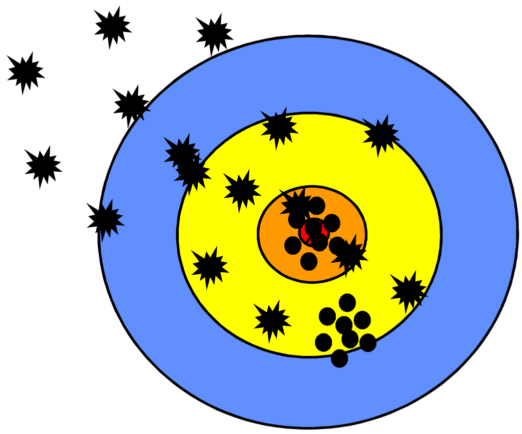
$$\Rightarrow \bar{x}_{1st} \neq \bar{x}_{2nd}$$

- **Modeling sample mean by random variables: Assuming iid $X_1, X_2, \ldots, X_n$**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Performance of Point Estimate and Bull's Eye Aiming

- Parameter to be estimated: bull's eye **Center**
- Two estimators: 57 Rifle and M16 Rifle



27

©Argon Chen

Unbiased Point Estimate

- A point estimator is itself a random variable with a distribution ⇒ Probability model of $\hat{\theta}$
- A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. If not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of θ
- Example: Is sample mean $\overline{X}$ the unbiased estimate of μ?

28

©Argon Chen

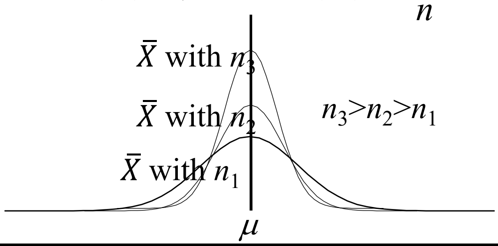
Is sample mean an Unbiased Estimator of mean?

- We model our observations $x_1, x_2,\ldots, x_n$ as independent and identically distributed (*iid*) $X_1,X_2,\ldots,X_n$ with μ and σ and sample mean can be modeled as a random variable:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Recall Mean and Variance of sample mean:

$$E(\bar{X}) = \mu \text{ and } V(\bar{X}) = \frac{\sigma^2}{n}$$



29

Is Sample Variance an Unbiased Estimator of Variance? E(S^2)=σ^2?

Assuming iid $X_1,X_2,\ldots,X_n$ with mean=μ and variance=σ^2

$$\mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2X_i\bar{X}+\sum_{i=1}^{n}\bar{X}^2\right]$$

$$=\frac{1}{n-1}\left\{\sum_{i=1}^{n}\mathrm{E}[X_i^2]-\mathrm{E}\left[2n\frac{\sum_{i=1}^{n}X_i}{n}\bar{X}-\sum_{i=1}^{n}\bar{X}^2\right]\right\}=\frac{1}{n-1}\left\{n\mathrm{E}[X_i^2]-\mathrm{E}\left[2n\bar{X}^2-n\bar{X}^2\right]\right\}$$

$$=\frac{1}{n-1}\left\{n[\sigma^2+\mu^2]-n\mathrm{E}\left[\bar{X}^2\right]\right\}=\frac{1}{n-1}\left\{n\sigma^2+n\mu^2-n\left[\frac{\sigma^2}{n}+\mu^2\right]\right\}$$

$$=\frac{1}{n-1}[(n-1)\sigma^2+n\mu^2-n\mu^2]=\sigma^2$$

30

Example: Point Estimate of Binomial Distribution Parameter

- Let x be the number of heads observed from n tosses of coin. What would be the most plausible estimate of p for the Binomial distribution model?

$$\hat{p} = \frac{x}{n}$$

- To evaluate the performance of $\hat{p}$, we model x to be a random variable X following a (n, p) Binomial distribution. Is $\hat{p}$ an unbiased estimate of p?

$$E(\hat{p}) = E(\frac{X}{n}) = \frac{E(X)}{n} = \frac{np}{n} = p$$

31 ©Argon Chen

Standard Error

- **Remember:** an estimator is itself a random variable with a distribution
- **Standard Error** of an estimator $\hat{\theta}$ is its S.D.

$$\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$$

- **Estimated Standard Error** is the estimate of $\sigma_{\hat{\theta}}$ often denoted by $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$
- Example: standard error of $\bar{X}$: $\sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$
- Example: Binomial experiment

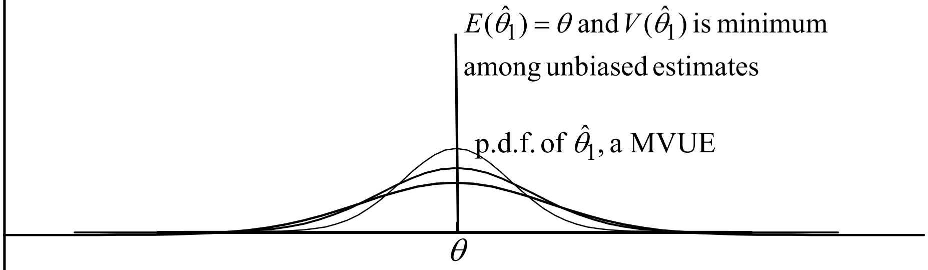
$$\sigma_{\hat{p}} = \sqrt{V(\hat{p} = X/n)} = \sqrt{\frac{V(X)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\hat{\sigma}_{\hat{p}} = s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(x/n)(1-x/n)}{n}} = \sqrt{\frac{x(n-x)}{n}}$$

32 ©Argon Chen

Performance of Point Estimate

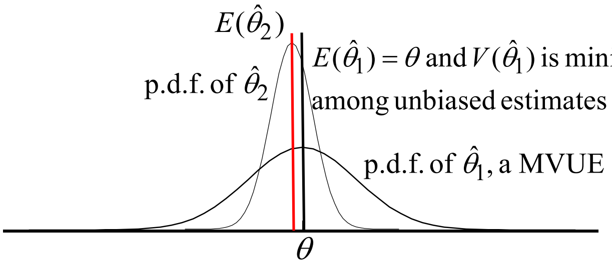
- How good is an estimate?
  - On target? Unbiased?
  - Very certain? Minimum variance?
- Minimum Variance Unbiased Estimator (MVUE)
  - Among all unbiased estimators, the one with the minimum variance
- Example: sample mean is a MVUE for normally distributed populations



33 ©Argon Chen

Preferable Point Estimate

- Is a MVUE the most preferable point estimate?



- There are different point estimates for the same model parameter
- Different models requires different estimates
- Different point estimates serve different needs

34 ©Argon Chen

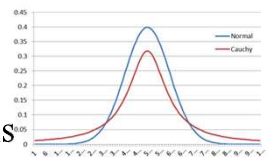
Example: Different Point Estimates of Mean

- Point estimates: $\overline{X}, \widetilde{X}, \overline{X}_e, \overline{X}_{tr(m)}$
- $\overline{X}$ is the arithmetic average called sample mean
- $\widetilde{X}$ is the median that is the center observation of the entire sample
- $\overline{X}_e$ is the extreme mean (an average of two extreme observations)
- $\overline{X}_{tr(m)}$ is a trimmed mean that trims m% of observations from each end of the sample

Different Mean Point Estimates for Different Distributions

- Point estimates: $\overline{X}, \widetilde{X}, \overline{X}_e, \overline{X}_{tr(m)}$
- $\overline{X}$ is the best for Normal distribution
- How about an estimator for Cauchy distribution:

$$f(x) = \frac{1}{\pi[1+(x-\mu)^2]}$$



Cauchy is bell-shaped with heavier tails

- $\overline{X}, \overline{X}_e$ are terrible since they are sensitive to outlying observations; $\widetilde{X}$ is quite good
- For Uniform distribution (no tails), $\overline{X}_e$ is the best
- $\overline{X}_{tr(m)}$ is not the best in all three situations, but it works reasonably well in all three!

Methods of Point Estimate

- Most often used methods:
  - moment estimator
  - maximum likelihood estimator (MLE)
- Where do you learn all these? Theories of Statistical Inference
- BUT don't worry! Most of reasonable estimators are from your intuitions
- Example: Binomial experiment

$$\hat{p} = X / n$$

37 ©Argon Chen

Moment Estimator

- Moments of distribution models:

  1^{st} moment=E(X), 2^{nd} moment=E(X^2),…

  m^{th} moment=E(X^m)
- Let $X_1, X_2,\ldots, X_n$ are ***independent*** random sample observations from a population following an ***identical*** probability model with p.m.f. or p.d.f. $f(X=x;\ \theta_1, \theta_2,\ldots, \theta_m)$. Then, the moment estimator of $\theta_1, \theta_2,\ldots, \theta_m$ are obtained by equating the first m sample moments to the corresponding first m model moments and solve for the estimators

38 ©Argon Chen

Example of Moment Estimator: Gamma Distribution

- To estimate the α and β of the Gamma distribution, equating the 1st and 2nd sample and model moments:

  sample mean=$\alpha\beta$=$E(X)$

  sample variance=$\alpha\beta^2$=$E(X^2)-E^2(X)$

  $\Rightarrow \hat{\beta}$=(sample variance)/(sample mean)

  $=S^2/\bar{X}$

  $\hat{\alpha}$=(sample mean)^2/(sample variance)

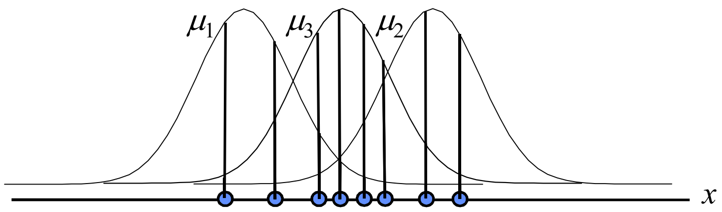
  $=\bar{X}^2/S^2$

Maximum Likelihood Estimate (MLE)

- The better the estimate of the parameter of a distribution model, the higher the value of the likelihood function substituted by the observed sample value.
- Let $X_1, X_2, \ldots, X_n$ are ***independent*** random sample observations from a population following an ***identical*** probability model with likelihood function $P(X)$ (discrete) or $f(X)$. Then, the joint joint likelihood for $X_1=x_1, X_2=x_2, \ldots, X_n=x_n$ is:

  $P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n)=P(X=x_1)P(X=x_2)\cdots P(X=x_n)$ or
  $f(X_1=x_1, X_2=x_2, \ldots, X_n=x_n)=f(X=x_1)f(X=x_2)\cdots f(X=x_n)$

- MLE is the estimate of a parameter that maximizes the joint likelihood function:



MLE for p of Binomial Distribution

- Let X follow a (n, p) binomial distribution and p is unknown.
  - we observe the number of successes x from n trials.
  - What is the MLE of p?

    $P(X{=}x){=}C^n_x p^x(1-p)^{n-x}$
- To maximize $P(x)$ w.r.t. p, take derivative of $P(x)$ w.r.t. p and set it to zero:

$$x\hat{p}^{x\text{-}1}(1\text{-}\hat{p})\text{-}(n\text{-}x)\hat{p}^{x}(1\text{-}\hat{p})^{n\text{-}x\text{-}1} = 0$$
$$\Rightarrow x(1\text{-}\hat{p}) = (n\text{-}x)\hat{p}$$
$$\Rightarrow \hat{p} = x/n$$

41 ©Argon Chen

MLE for μ of Normal Distribution

- Let X follow a (μ, σ^2) normal distribution and μ is unknown.
  - we take a sample of n observed values $x_1, x_2, \ldots, x_n$.
  - the joint likelihood function

$$f(x_1,\ldots,x_n \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right] = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-\sum(x_i-\mu)^2}{2\sigma^2}\right]$$

- Maximizing $f(x_1,\ldots,x_n)$ is equivalent to maximizing $\log f(x_1,\ldots,x_n)$. Take derivative of $\log f(x_1,\ldots,x_n)$ and set it to zero:

$$\frac{d}{d\mu}\log f(x_1,\ldots,x_n \mid \mu) = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

42 ©Argon Chen

Which Probability Model Fits Best?

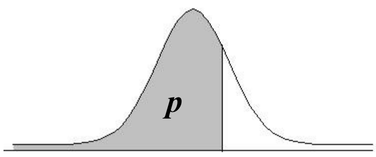
- Sample observations can be used to estimate parameters of any given probability model.
  - MLE can be used to estimate p of Geometric distribution as well as λ of Poisson distribution
- Which probability distribution model fits best to the sample observed data?
  - Goodness-of-Fit Tests
  - **Q-Q (P-P) plot: an effective visualized goodness of fit**

43 ©Argon Chen

Percentile and Sample Percentile

- Cumulative probability $F(\eta(p))=p$

  $\eta(p)$ is called the **(100*p*)th percentile (Quantile) = $F^{-1}(p)$**



$\eta(p)$

- Sort the n sample observations from the smallest to the largest and the ***sample cumulative probability*** of the ith smallest observation=***100(i-0.5)/n***
- The ith smallest observation = **[*100*(*i-0.5*)/*n*]th sample percentile is an estimate of [*100*(*i-0.5*)/*n*]th percentile**

44 ©Argon Chen

Probability Plot (Q-Q Plot)

Step 1: Sort the data from the smallest to the largest

Step 2: Calculate the sample cumulative probabilities =***100(i-0.5)/n***

Step 3: Assume a probability distribution model (***F***) and estimate probability distribution parameters

Step 4: Calculate the percentiles of the sample cumulative probabilities = ***F***$\boldsymbol{F^{-1}((i-0.5)/n)}$

Step 5: Plot $\begin{pmatrix} [100(i-.5)/n]\text{th percentile,} & i\text{th smallest} \\ \text{of the distribution} & \text{sample observation} \end{pmatrix}$

on the X-Y plane. If the observations follow the assumed distribution, the points form roughly a 45^o line

45 ©Argon Chen

Example: 350Å SiO2 Assumed to Follow Normal Distribution

Step 1: Sort the thickness data from 345.6 to 355.6 (X_i)

Step 2: Calculate the sample cumulative probabilities

$$\hat{p}_i = (i - 0.5)/n$$

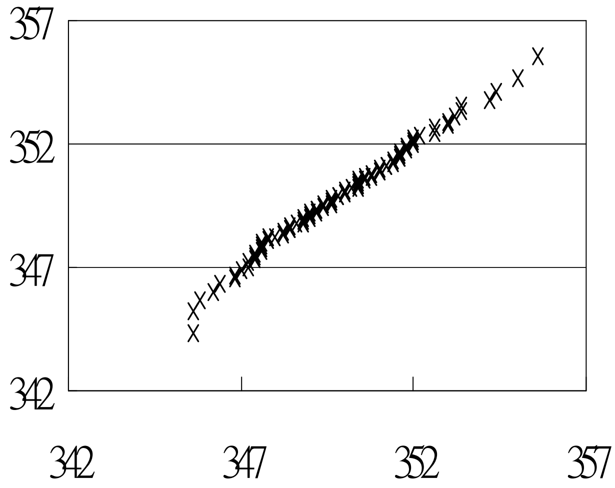
Step 3: Assume a **normal** probability distribution and estimated mean=349.91 and sample s.d.=2.235

Step 4: Calculate the *percentiles* of the sample cumulative probabilities $N^{-1}(\hat{p}_i;349.91, 2.235)$

Step 5: Plot $[N^{-1}(\hat{p}_i;349.91, 2.235), X_i]$

46 ©Argon Chen

Example: 350Å SiO2 Assumed to Follow Normal Distribution



47 ©Argon Chen

Normal Probability Plot

- $X \sim N(\mu, \sigma)$
- $Z=(X-\mu)/\sigma \sim N(0, 1)$

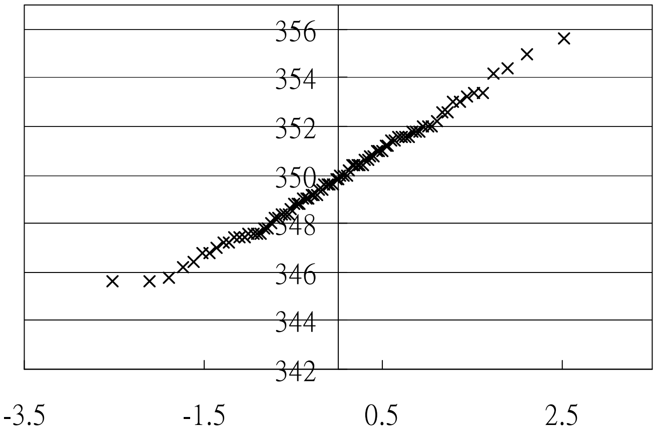
([100(i-.5)/n]th z percentile, ith smallest obs. x)

$\Rightarrow$ form a line: $X=\sigma Z+\mu$

is a line with slop σ and intercept μ

48 ©Argon Chen

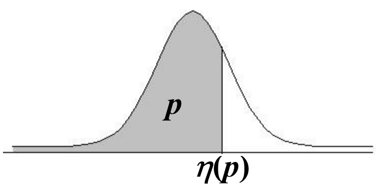
Example: 350Å SiO2 Normal Probability Plot



49 ©Argon Chen

Percentile and Sample Percentile

- Cumulative probability $F(\eta(p))=p$
  $\eta(p)$ is called the **(100*p*)th percentile (Quantile) = $F^{-1}(p)$**



- Sort the *n* sample observations from the smallest to the largest and the ***sample cumulative probability*** of the *i*th smallest observation=***100(i-0.5)/n***
- The *i*th smallest observation = **[*100*(*i-0.5*)/*n*]th sample percentile is an estimate of [*100*(*i-0.5*)/*n*]th percentile**

50 ©Argon Chen

Probability Plot (Q-Q Plot)

Step 1: Sort the data from the smallest to the largest

Step 2: Calculate the sample cumulative probabilities =***100(i-0.5)/n***

Step 3: Assume a probability distribution model (***F***) and estimate probability distribution parameters

Step 4: Calculate the percentiles of the sample cumulative probabilities = $\boldsymbol{F^{-1}(100(i-0.5)/n)}$

Step 5: Plot

$$\begin{pmatrix} [100(i-.5)/n]\text{th percentile,} & i\text{th smallest} \\ \text{of the distribution} & \text{sample observation} \end{pmatrix}$$

on the X-Y plane. If the observations follow the assumed distribution, the points form roughly a 45° line

51 ©Argon Chen

Example: 350Å SiO2 Assumed to Follow Normal Distribution

Step 1: Sort the thickness data from 345.6 to 355.6 (X_i)

Step 2: Calculate the sample cumulative probabilities ($\hat{p}_i$)

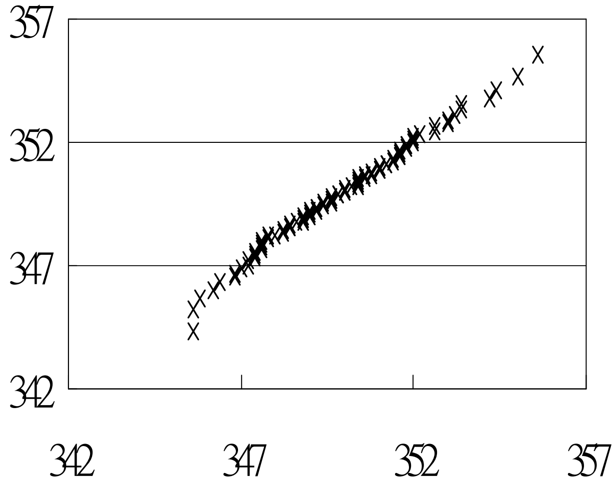
Step 3: Assume a **normal** probability distribution and estimated mean=349.91 and sample s.d.=2.235

Step 4: Calculate the *percentiles* of the sample cumulative probabilities $N^{-1}(\hat{p}_i;349.91, 2.235)$

Step 5: Plot $[N^{-1}(\hat{p}_i;349.91, 2.235), X_i]$

52 ©Argon Chen

Example: 350Å SiO2 Assumed to Follow Normal Distribution



53 ©Argon Chen

Normal Probability Plot

- $X \sim N(\mu, \sigma)$
- $Z=(X-\mu)/\sigma \sim N(0, 1)$

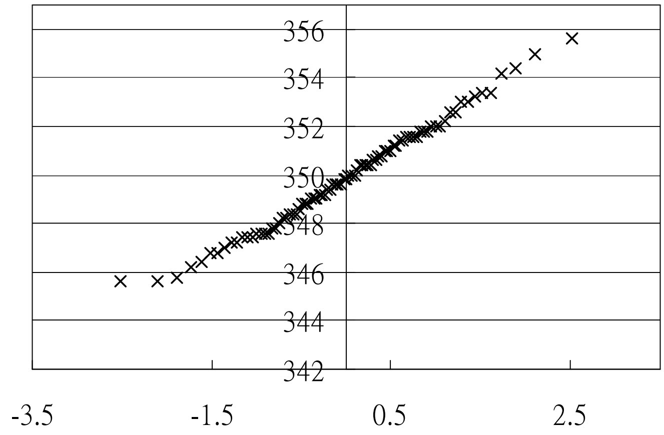
([100(i-.5)/n]th z percentile, ith smallest obs. x)

$\Rightarrow$ form a line: $X=\sigma Z+\mu$

is a line with slop σ and intercept μ

54 ©Argon Chen

Example: 350Å SiO2 Normal Probability Plot

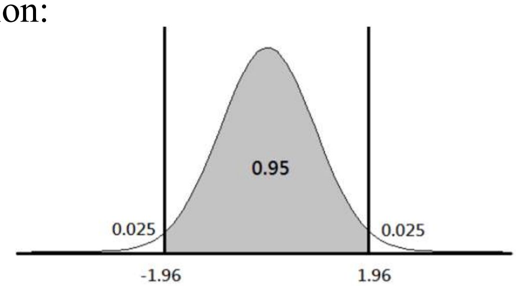


55 ©Argon Chen

Probability Interval of Normal Sample Mean

- Actual sample observations $x_1, x_2, \ldots, x_n$ from random sample $X_1, X_2, \ldots, X_n$ following $N(\mu, \sigma)$. Then, $\bar{X}$ follows $N(\mu, \sigma/\sqrt{n})$
- Standardization:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$



$$\Rightarrow P(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = .95$$

56 ©Argon Chen

Confidence Interval (C.I.) of Normal Mean

$$P(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = .95$$

$$\Rightarrow P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = .95$$

- What does this mean?
- Remember that the sample mean is an estimator of the mean and is itself a ***random variable*** with uncertainty.
- With confidence interval, you are about 95% sure that the true mean will be in the range of

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

- What if I would like to be **99%** sure?

57 ©Argon Chen

100(1-α)% Confidence Interval

- a **100(1-α)%** confidence interval for the mean μ of a normal population when the value of σ is known is given by

$$(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

- Example: 99%, $Z_{\alpha/2}$=?
- In reality, what is the difficulty? σ is usually unknown!

58 ©Argon Chen

Large-sample Confidence Interval

- If sample size n is sufficiently large

$$Z \cong \frac{\overline{X}-\mu}{S/\sqrt{n}} \quad (S\text{: sample S.D.})$$

- Z is then approximately $N(0, 1)$

$$(\bar{x}+z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x}+z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}})$$

is a **large-sample confidence interval** for μ with confidence level approximately 100(1-α)%

59 ©Argon Chen

Example: 350Å SiO_2 Thickness

- Sample size n=85
- Sample mean=349.91 and sample s.d.=2.235
- 95% confidence interval of the thickness mean?
- 95%=100(1-0.05)% ⇒ α=0.05
- Thickness mean 95% confidence interval:

$$(349.91+z_{0.025}\frac{2.235}{\sqrt{85}}, 349.91+z_{0.975}\frac{2.235}{\sqrt{85}}) =$$
$$(349.91-1.96\cdot 0.2424, 349.91+1.96\cdot 0.2424) =$$
$$(349.435, 350.385)$$

- What if n is considered small or the sample s.d. is not considered reliable?

60 ©Argon Chen

t statistics: C.I. Interval for Unknown σ

- $\overline{X}$ is the average of a random sample of size n from a normal distribution with mean μ, Then, the random variable

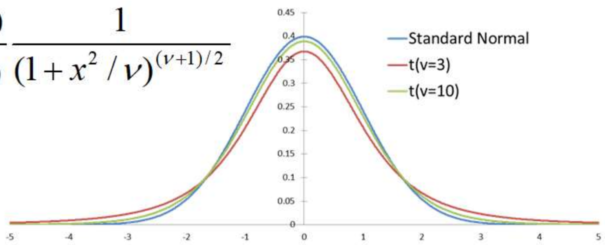
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a probability distribution called a ***t*** distribution with $n-1$ degrees of freedom

61 ©Argon Chen

t-Distribution

$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \frac{1}{(1+x^2/\nu)^{(\nu+1)/2}}$$

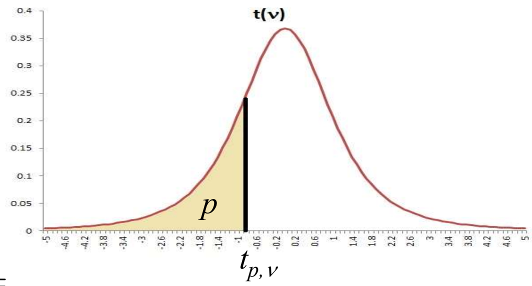


- Bell-shaped & centered at 0
- $\nu\uparrow$ distribution spread $\downarrow$
- The distribution spreads wider than the normal distribution (heavier tails)
- $\nu\to\infty$ $t_\nu\to$standard normal $N(0, 1)$

62 ©Argon Chen

Confidence Interval using *t*-Statistic

- Let $t_{p,\nu}$ denotes



$$\Rightarrow \quad P(t_{\alpha/2,\nu} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{1-\alpha/2,\nu}) = 1-\alpha$$

Then a 100(1-α)% confidence interval for μ is: $(\bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}})$

 ©Argon Chen

Example: 350Å SiO_2 Thickness

- Sample size n=85
- Sample mean=349.91 and sample s.d.=2.235
- 95% confidence interval of the thickness mean?
- 95%=100(1-0.05)% ⇒ α=0.05
- Thickness mean 95% confidence interval using Z:

$$(349.91 + z_{0.025}\frac{2.235}{\sqrt{85}},\ 349.91 + z_{0.975}\frac{2.235}{\sqrt{85}}) =$$

$$(349.91 - 1.96\cdot 0.2424,\ 349.91 + 1.96\cdot 0.2424) = (349.435,\ 350.385)$$

- Thickness mean 95% confidence interval using t:

$$(349.91 + t_{0.025,84}\frac{2.235}{\sqrt{85}},\ 349.91 + t_{0.975,84}\frac{2.235}{\sqrt{85}}) =$$

$$(349.91 - 1.9886\cdot 0.2424,\ 349.91 + 1.9886\cdot 0.2424) = (349.428,\ 350.392)$$

 ©Argon Chen

Tests of Hypotheses

- The motivation: to reject an initial claim and to statistically prove that a scientific effort really makes differences
- Example: Medical experiment
- Initial claim

  Null hypothesis H_0
- Claim otherwise

  Alternative hypothesis H_1

65 ©Argon Chen

Errors in Hypothesis Tests

- Type I error: rejecting the null hypothesis H_0 when it is true
- Type II error: not rejecting H_0 when H_0 is false
- Probability of type I error (α)
- Probability of type II error (β)

66 ©Argon Chen

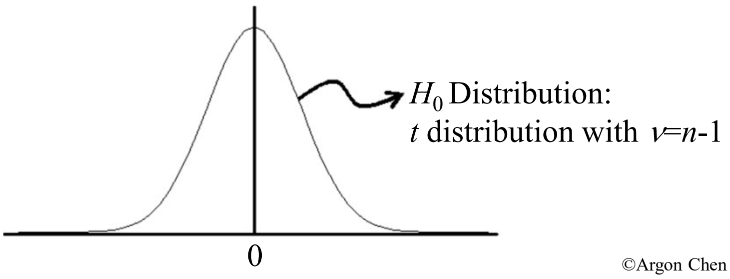
Test Statistic and Distribution under H_0 (Example: Testing Normal Mean)

$$H_0: \mu = \mu_0$$

Test statistic: $t\text{-test} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Distribution under H_0: **t-Distribution with ν=n-1**

μ=$\mu_0 \Rightarrow E(t)$=0



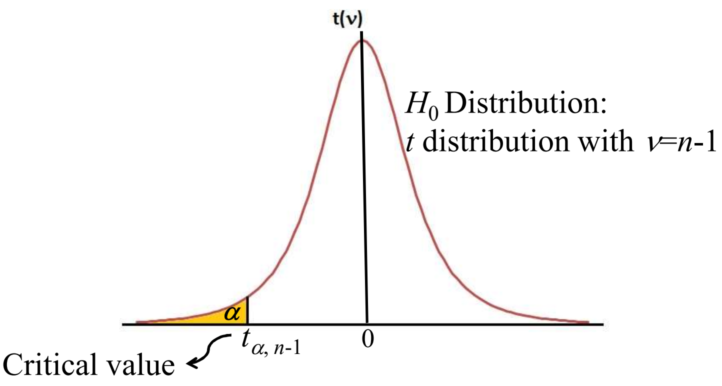
67 ©Argon Chen

Criteria to Reject H_0: Reject Region

$H_0: \mu = \mu_0$ Test statistic: $t\text{-test} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Distribution under H_0: t-distribution

Reject H_0 when $t\text{-test} \le t_{\alpha, n\text{-}1} \Rightarrow H_1: \mu < \mu_0$

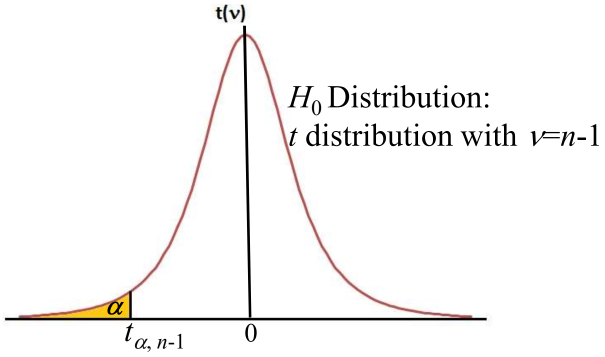


68 ©Argon Chen

Type I Error Probability of Rejecting the Null Hypothesis

Reject H_0 when t-test $\leq t_{1-\alpha,\, n-1} \Rightarrow H_1\colon \mu < \mu_0$

Probability ($\mu = \mu_0$ but you reject H_0 and accept H_1) $= \alpha$



Hypothesis Test Procedure

1 Choose a test statistic: a function of the sample data with a known probability distribution model under $\boldsymbol{H_0}$

2 Choose the Type I error probability (significance level) α to find the reject region and critical value based on the distribution of the test statistic under $\boldsymbol{H_0}$

3 The $\boldsymbol{H_0}$ will then be rejected if and only if the observed or computed test statistic values falls in the reject region

Test about Population Mean

H_0: $\mu = \mu_0$

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

H_1	Reject Region	
$\mu > \mu_0$	$t \geq t_{1-\alpha, n-1}$	one-tailed test
$\mu < \mu_0$	$t \leq t_{\alpha, n-1}$	
$\mu \neq \mu_0$	$t \geq t_{1-\alpha/2, n-1}$ or $t \leq t_{\alpha/2, n-1}$	two-tailed test

71 ©Argon Chen

Another View of Reject Region

Let H_1 be $\mu \neq \mu_0$ then reject H_0 when $t \leq t_{\alpha/2, n-1}$ or $t \geq t_{1-\alpha/2, n-1}$

$$\Rightarrow \quad \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \leq t_{\alpha/2, n-1} \text{ or } \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \geq t_{1-\alpha/2, n-1}$$

$$\mu_0 \leq \bar{x} - t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}} \text{ or } \mu_0 \geq \bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

$$\mu_0 \leq \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \text{ or } \mu_0 \geq \bar{x} + t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

$\Rightarrow$ when μ_0 is outside the 100(1-α)% confidence interval of mean estimated by $\bar{x}$:

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

we reject it and accept $\mu \neq \mu_0$ with Type I error probability = α

72 ©Argon Chen

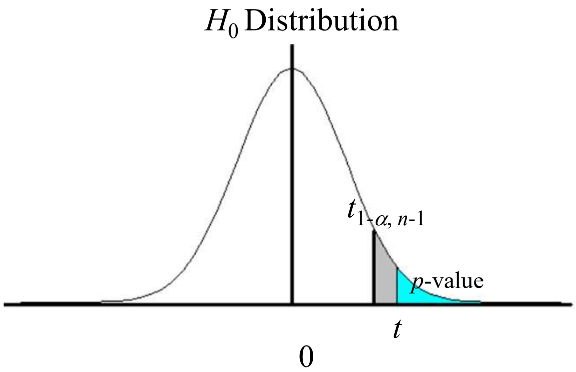
p-value in Hypothesis Tests

- Another way of presenting the test result
- The **p-value is the smallest level of significance at which H_0 would be rejected** when a specified test procedure is used on a given data set. Once the p-value has been determined, the conclusion at any particular level α results from comparing the p-value to α:

73 ©Argon Chen

p-value in one-tailed Hypothesis Tests

- **p-value for one-tailed test:** (example:H_1:$\mu>\mu_0$)



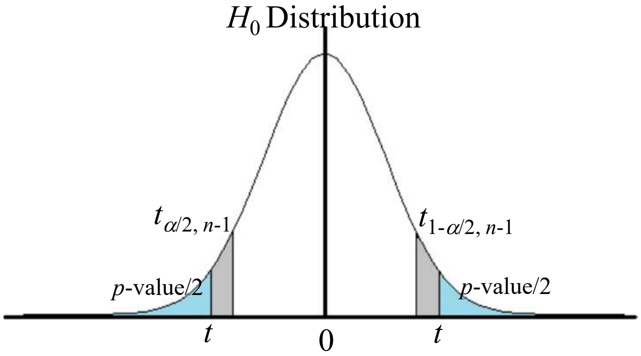
a. $t > t_{1-\alpha,n-1} \Rightarrow$ P(t-test $\geq t$)= p-value $\leq \alpha \Rightarrow$Reject H_0

b. p-value $> \alpha \Rightarrow$Accept H_0

74 ©Argon Chen

p-value in two-tailed Hypothesis Tests

- ***p*-value for two-tailed test:** (example:H_1:$\mu \neq \mu_0$)



a. $t \geq t_{1-\alpha/2, n-1} \Rightarrow$P($t$-test $\geq t$)$\leq \alpha/2 \Rightarrow$2P($t$-test $\geq t$)=p-value $\leq \alpha$
or $t \leq t_{\alpha/2, n-1} \Rightarrow$P($t$-test $\leq t$)$\leq \alpha/2 \Rightarrow$2P($t$-test $\leq t$)=p-value $\leq \alpha$
$\Rightarrow$ Reject H_0

75 b. p-value > $\alpha \Rightarrow$Accept H_0 ©Argon Chen

Testing 350Å SiO_2 Thickness Mean

- Null hypothesis: Thickness Mean= Target
  H_0: μ=350
- Sample size n=85
- $\bar{x}$ =349.91 and s=2.235 $\Rightarrow H_1$: μ < 350
- Test statistic= $\dfrac{349.91-350}{2.235/\sqrt{85}} = -0.37$
- Type I error prob.=0.05 $\Rightarrow$ α=0. 05
- Reject μ=350 if t-test $\leq t_{0.05,84}$=–1.663
  but –0.37>–1.663 $\Rightarrow \mu$=350 can't be rejected
- p-value=0.356157 > 0.05 $\Rightarrow$do not reject H_0

76 ©Argon Chen

Testing A Pair of Means

- Testing problem: are the means the same?
  - That is: H_0: $\mu_x = \mu_y \Rightarrow \mu_x - \mu_y = 0$
- Choice of the test statistic:
  - Compare two sample means $\bar{X}$ with sample size m and $\bar{Y}$ with sample size n: $\bar{X} - \bar{Y}$
  - Question: what would be the distribution of $\bar{X} - \bar{Y}$?
  - Assuming X and Y are normally distributed with the same mean and a common variance σ^2, then $\bar{X} - \bar{Y}$ follows a normal distribution with mean=$\mu_x - \mu_y$ and variance=$\sigma^2/m + \sigma^2/n = \sigma^2(\frac{1}{m} + \frac{1}{n})$
  - S.D. of $\bar{X} - \bar{Y}$: $\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}$, how to estimate σ?

77 ©Argon Chen

Pooled Estimator of σ^2

- The pooled estimator of the common variance σ^2, denoted by S_p^2, is the weighted average of the individual sample variances **weighted by the degree of freedom, i.e., *m*-1 and *n*-1**:

$$S_p^2 = \frac{m-1}{m-1+n-1} S_x^2 + \frac{n-1}{m+n-2} S_y^2$$

- Estimate of $\bar{X} - \bar{Y}$ S.D.: $S_p\sqrt{\frac{1}{m} + \frac{1}{n}}$

78 ©Argon Chen

Testing Mean Difference

H_0: $\mu_x - \mu_y = \Delta$ (Δ=0 for testing the same means)

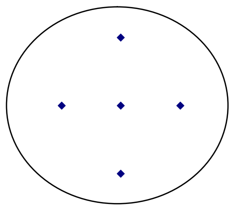
Test statistic: $$t = \frac{(\bar{x} - \bar{y}) - \Delta}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

H_1	Reject Region	
$\mu_x - \mu_y > \Delta$	$t \geq t_{1-\alpha, m+n-2}$	one-tailed test
$\mu_x - \mu_y < \Delta$	$t \leq t_{\alpha, m+n-2}$	
$\mu_x - \mu_y \neq \Delta$	$t \geq t_{1-\alpha/2, m+n-2}$ or $t \leq t_{\alpha/2, m+n-2}$	two-tailed test

79 ©Argon Chen

Testing 350Å SiO_2 Means at Different Sites

Center	Top	Bottom	Left	Right
349	349	352	347	353
349	347	352	349	353
348	346	351	346	351
352	350	349	354	353
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.



- Null hypothesis: H_0: $\mu_{center} = \mu_{top}$
- Sample size $m=n=86$
- $\bar{x}_{center}$ =349.09 and s^2_{center}=5.12
  $\bar{x}_{top}$ =348.48 and s^2_{top}=5.52
- Pooled estimate of σ: $s_p = \sqrt{\frac{85 s^2_{center} + 85 s^2_{top}}{86+86-2}}$ =2.31
- Test statistic = $\frac{349.09 - 348.48}{2.31\sqrt{\frac{1}{86} + \frac{1}{86}}}$ = 1.752
- Type I error prob.=0.05 ⇒ two-tailed α/2=0. 025; 1-α/2=0. 975
- Do not reject $\mu_{center} = \mu_{top}$ as t-test =1.752 < $t_{0.975, 86+86-2}$= 1.974
  ⇒ do not accept $\mu_{center} \neq \mu_{top}$
- p-value=2× Prob($t_{86+86-2}$ >1.752)=0.0816 > 0.05 ⇒ do not reject H_0

80 ©Argon Chen

Testing Variance

- Testing problem: is the variance unchanged?
  - That is: H_0: $\sigma^2 = \sigma_0^2$
- Choice of the test statistic:
  - Sample variance S^2 seems to be a good test statistics since it's an unbiased estimate of variance
  - Question: what would be the distribution of S^2 under H_0 ($\sigma^2 = \sigma_0^2$)?

81 ©Argon Chen

Distribution of Sample Variance

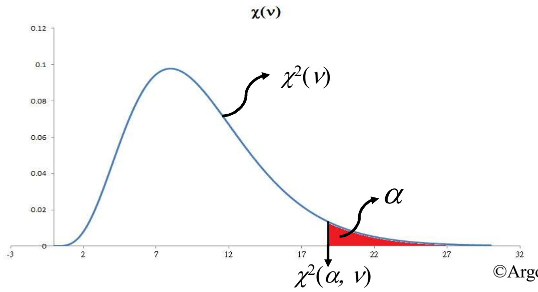
- To establish a hypothesis, one must know the assumed distribution model behind the sample. What would be the distribution model behind the variance?
- Recall: If $X_i \sim N(0, 1)$ then $\Sigma_n X_i^2 \sim \chi^2(n)$
- Recall: $(X-\mu)/\sigma \sim N(0, 1)$
- Therefore: $\Sigma_n[(X_i-\mu)/\sigma]^2 = \Sigma_n(X_i-\boxed{\mu})^2/\sigma^2 \sim \chi^2(\boxed{n})$
- Recall: $S^2 = \Sigma_n(X_i-\bar{X})^2/(n-1)$
- Result: $(n-1)\, S^2/\sigma^2 = \Sigma_n(X_i-\boxed{\bar{X}})^2/\sigma^2 \sim \chi^2(n\boxed{-1})$
- If $X_1, X_2, \ldots, X_n$ is a sample from a normal population with mean μ and variance σ^2, then $(n-1)\, S^2/\sigma^2$ follows a χ^2 distribution with $\nu = n-1$.

©Argon Chen

82

Testing H_0: $\sigma^2 = \sigma_0^2$

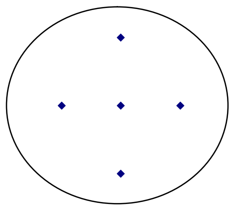
- Test statistic: $(n-1)S^2/\sigma_0^2$
- Under H_0 ($\sigma^2 = \sigma_0^2$), $(n-1)S^2/\sigma_0^2$ follows $\chi^2(n-1)$
- Reject region: determine critical value $\chi^2(\alpha, \nu)$ (CHISQ.INV.RT(α, ν) in Excel) and reject region based on α
- H_1: $\sigma^2 > \sigma_0^2$



Testing 350Å SiO_2 Thickness Variance

- Null hypothesis: H_0: $\sigma^2_{top} = 5.0$
- Sample size n=86
- $\bar{x}_{top}$ =348.48 and s^2_{top}=5.52
- Test statistic = $(86-1)\frac{5.52}{5.0} = 93.89$
- Type I error prob. α=0.05
- $\chi^2(0.05, 86-1) = 107.52$
- Do not reject $\sigma^2_{top} = 5.0$ as test-stat =93.89 < 107.52 $\Rightarrow$ do not accept $\sigma^2_{top} > 5.0$
- p-value= $\text{Prob}(\chi^2_{85} > 93.89) = 0.238 > 0.05 \Rightarrow$ do not reject H_0

Center	Top	Bottom	Left	Right
349	349	352	347	353
349	347	352	349	353
348	346	351	346	351
352	350	349	354	353
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.



Testing A Pair of Variances

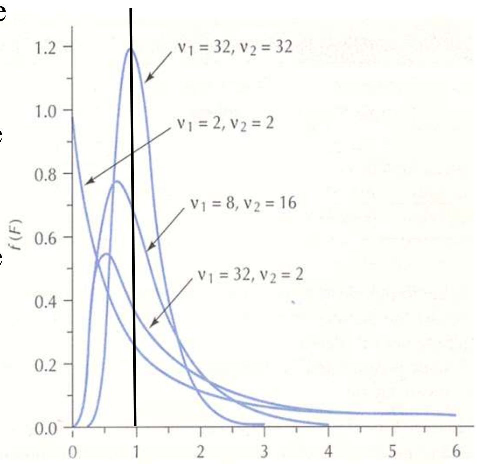
- Testing problem: are the variances the same?
  - That is: H_0: $\sigma_x^2 = \sigma_y^2$
- Choice of the test statistic:
  - Compare two sample variances S_x^2 with sample size n_1 and S_y^2 with sample size n_2:

$$S_x^2/S_y^2$$

  - Question: what would be the distribution of S_x^2/S_y^2 under H_0 ($\sigma_x^2 = \sigma_y^2$)?

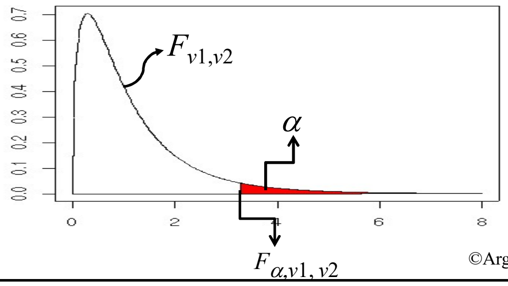
Distribution of Sample Variance Ratio

- Two ***normal*** populations have the ***same variance***
- S_x^2: sample variances from population 1 with sample size n_1
- S_y^2: sample variances from population 2 with sample size n_2
- Then, S_x^2/S_y^2 follows $F_{v1,\,v2}$ distribution with $v_1=n_1-1$; $v_2=n_2-1$
- $E(S_x^2/S_y^2) = v_2/(v_2-2)$ if $v_2>2$
- $E(S_x^2/S_y^2) \to 1$ as $v_2 \to$ large number



Testing H$_0$: $\sigma_x^2 = \sigma_y^2$

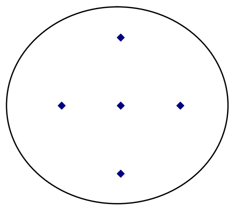
- Test statistic: S_x^2/S_y^2
- Under H$_0$ ($\sigma_x^2 = \sigma_y^2$), S_x^2/S_y^2 follows $F_{v1,v2}$
- Reject region: determine critical value $F_{\alpha,v1,v2}$ (F.INV.RT(α, ν1, ν2) in Excel) and reject region based on α
- H_1: $\sigma_x^2 > \sigma_y^2$



Testing Variances at Different Sites

- Null hypothesis:
  H_0: $\sigma^2_{bottom} = \sigma^2_{center}$
- Sample size $m=n=86$
- $s^2_{bottom}=7.88$
  $s^2_{center}=5.12$
- F-test statistic $= \frac{7.88}{5.12} = 1.538$
- Type I error prob. α=0.05
- $F_{0.05,\,86\text{-}1,86\text{-}1}=1.432$
- Reject $\sigma^2_{bottom} = \sigma^2_{center}$ as F-test $=1.538 > 1.432$
  $\Rightarrow$ accept $\sigma^2_{bottom} > \sigma^2_{center}$
- p-value=Prob($F_{86\text{-}1,\,86\text{-}1}$ >1.538)=0.024<0.05 $\Rightarrow$ Reject H$_0$

Center	Top	Bottom	Left	Right
349	349	352	347	353
349	347	352	349	353
348	346	351	346	351
352	350	349	354	353
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.



Testing a Proportion

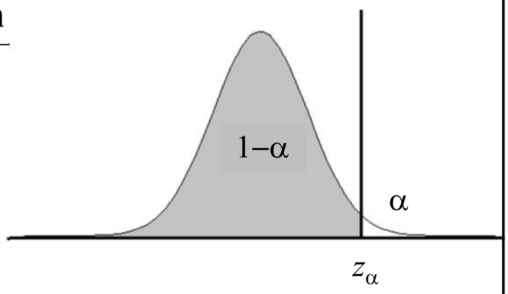
- Testing problem: is the proportion of occurrences equal to p_0?
  - That is: H_0: $p=p_0$
- Choice of the test statistic:
  - $\boldsymbol{X}$: number of occurrences in $\boldsymbol{n}$ trials
  - $\boldsymbol{X}$ follows Binomial distribution $b(x; \boldsymbol{p_0}, \boldsymbol{n})$ under null hypothesis
- Given significance level (type I error prob.) α and H_1: $p>p_0$
  - Reject region: $x>k^*$ where k^* is the smallest value of k for which $\sum_{i=k}^{n} C_i^n p_0^i (1-p_0)^{n-i} \le \alpha$.

89 ©Argon Chen

Testing the Proportion with Large *n*

- H_0: $p=p_0$
- Test statistic: $\boldsymbol{Z}=\dfrac{\boldsymbol{X-np_0}}{\sqrt{\boldsymbol{np_0(1-p_0)}}}$ ~ standard normal with a large $\boldsymbol{n}$ (recall: $E(X)=np$, $Var(X)=np(1-p)$) under null hypothesis

H_1	Reject Region
$\boldsymbol{p > p_0}$	$\boldsymbol{z \ge z_\alpha}$
$\boldsymbol{p < p_0}$	$\boldsymbol{z \le -z_\alpha}$
$\boldsymbol{p \ne p_0}$	$\boldsymbol{z \ge z_{\alpha/2}}$ or $\boldsymbol{z \le -z_{\alpha/2}}$



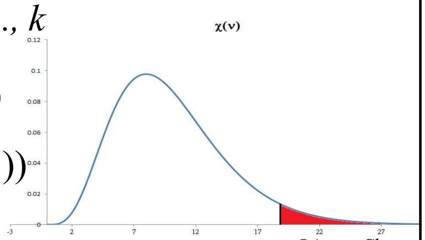
©Argon Chen

Testing Proportions

- Testing problem: do the occurrences of different classes agree with the hypothesis that the occurrence probability of class i (for i=1,…., k) equals to p_i and $\Sigma p_i=1$
- Let observed occurrences of class i be X_i (for i=1,…., k) in a total of n observations ($\Sigma X_i=n$)
- Test statistic: $C^2=\sum_{i=1}^{k}\frac{(X_i-np_i)^2}{np_i}$ follows $\chi^2(\nu=k\text{-}1)$ under null hypothesis
- Reject H_0: P(Class=i)=p_i i=1,…., k

  if $c^2=\sum_{i=1}^{k}\frac{(x_i-np_i)^2}{np_i}\geq\chi^2(\alpha, k-1)$

(In Excel, CHISQ.INV.RT(α, $k-1$))

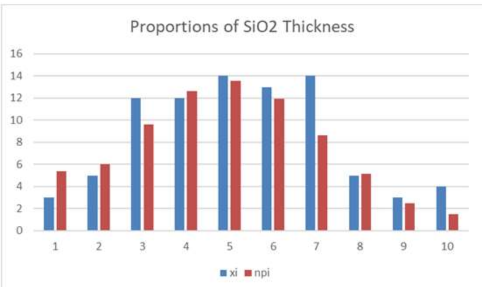


91 ©Argon Chen

Testing the Frequency Distribution

- Example: Is the frequency distribution for 85 readings of SiO_2 average fit the proportions based on Normal distribution N(349.91, 2.235)?

	Class Interval	x_i	Relative Freq.	np_i
1	~346.5	3	0.0353	5.40
2	346.6~347.5	5	0.0588	6.05
3	347.6~348.5	12	0.1412	9.64
4	348.6~349.5	12	0.1412	12.61
5	349.6~350.5	14	0.1647	13.54
6	350.6~351.5	13	0.1529	11.93
7	351.6~352.5	14	0.1647	8.63
8	352.6~353.5	5	0.0588	5.13
9	353.6~354.5	3	0.0353	2.50
10	354.6~	4	0.0471	1.53



92 ©Argon Chen

Testing the Proportions

- H_0: P(Class=i)=p_i i=1,…., 10

	Class Interval	x_i	Relative Freq.	np_i	$(x_i-np_i)^2/np_i$
1	~346.5	3	0.0353	5.40	1.067
2	346.6~347.5	5	0.0588	6.05	0.181
3	347.6~348.5	12	0.1412	9.64	0.580
4	348.6~349.5	12	0.1412	12.61	0.029
5	349.6~350.5	14	0.1647	13.54	0.016
6	350.6~351.5	13	0.1529	11.93	0.096
7	351.6~352.5	14	0.1647	8.63	3.340
8	352.6~353.5	5	0.0588	5.13	0.003
9	353.6~354.5	3	0.0353	2.50	0.101
10	354.6~	4	0.0471	1.53	4.008
	Total	85		c^2	9.421
				p-value	0.399
				χ^2(0.05, 10-1)	16.919

- c^2=9.421< χ^2(0.05, 10−1)=16.919
- Do not reject the H_0: the SiO_2 thickness proportions are the same as the normal distribution proportions

93 ©Argon Chen