UCLA STAT 251

Statistical Methods for the Life and Health Sciences

- **Instructor:** **Ivo Dinov,**

  Asst. Prof. In Statistics and Neurology

  University of California, Los Angeles, Winter 2002

  http://www.stat.ucla.edu/~dinov/

Tools for Exploring Univariate Data

- Types of variables
- Presentation of data
- Simple plots
- Numerical summaries
- Repeated and grouped data
- Qualitative variables

TABLE 2.1.1 Data on Male Heart Attack Patients

A subset of the data collected at a Hospital is summarized in this table. Each patient has measurements recorded for a number of variables – ID, Ejection factor (ventricular output), blood systolic/diastolic pressure, etc.

- Reading the table
- Which of the measured variables (age, ejection etc.) are useful in predicting how long the patient may live.
- Are there relationships between these predictors?
- variability & noise in the observations hide the message of the data.

TABLE 2.1.1 Data on Male Heart Attack Patients

ID	EJEC	SYS-VOL	DIA-VOL	OCCLU	STEN	TIME	OUT-COME	AGE	SMOKE	BETA	CHOL[a]	SURG
390	72	36	131	0	0	143	0	49	2	2	59	0
279	52	74	155	37	63	143	0	54	2	2	68	1
288	59	39	94	0	0	135	0	47	1	2	63	0
407	67	39	117	0	73	53	1	57	2	2	62	2

[a] NA = Not Available (missing data code).

TABLE 2.1.1 Data on Male Heart Attack Patients

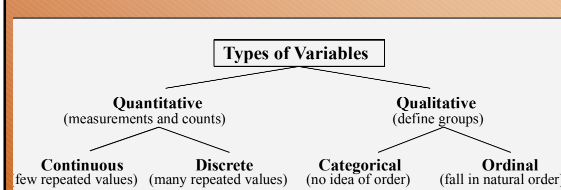
ID	EJEC	SYS-VOL	DIA-VOL	OCCLU	STEN
390	72	36	131	0	0
279	52	74	155	37	63
391	62	52	137	33	47
201	50	165	329	33	30
202	50	47	95	0	100
69	27	124	170	77	23
310	60	86	215	7	50
392	72	37	132	40	10
311	60	65	163	0	40
288	59	39	94	0	0
407	67	39	117	0	73

[a] NA = Not Available (missing data code).

Types of variable

- ***Quantitative*** variables are ***measurements*** and counts
  - Variables with ***few*** *repeated values* are treated as ***continuous***.
  - Variables with ***many*** *repeated values* are treated as ***discrete***
- ***Qualitative*** variables (a.k.a. factors or class-variables) describe ***group membership***

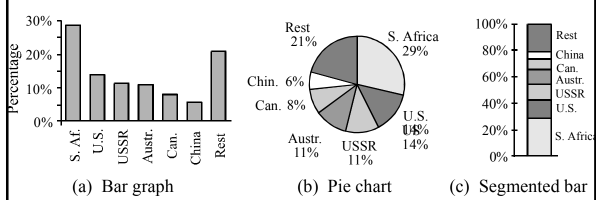
Distinguishing between types of variable



Questions …

- What is the difference between quantitative and qualitative variables?
- What is the difference between a discrete variable and a continuous variable?
- Name two ways in which observations on qualitative variables can be stored on a computer. (strings/indexes)
- When would you treat a discrete random variable as though it were a continuous random variable?
  - Can you give an example? ($34.45, bill)

Different graphs of the same set of numbers – percentages of the world's gold production in 1991



Questions …

- For what two purposes are tables of numbers presented? (convey information about trends in the data, detailed analysis)
- When should you round numbers, and when should you preserve full accuracy?
- How should you arrange the numbers you are most interested in comparing? (Arrange numbers you want to compare in columns, not rows. Provide written/verbal summaries/footnotes. Show row/column averages.)
- Should a table be left to tell its own story?

The dot plot

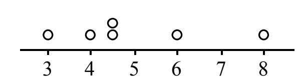


Figure 2.3.1 Dot plot.

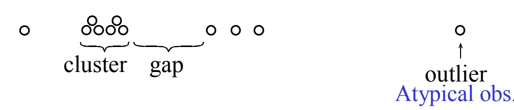


Figure 2.3.2 Dot plot showing special features.

Example of exploiting gaps and clusters

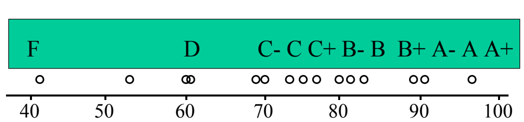


Figure 2.3.3 Grading of a university course.

Scale breaks

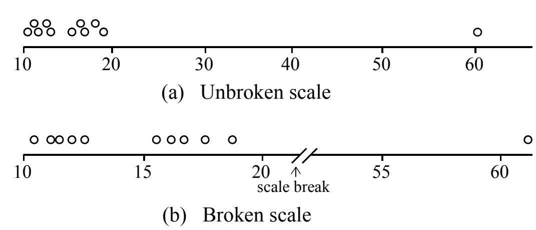


Figure 2.3.4 Dot plot with and without a scale break.

A labeled dot plot

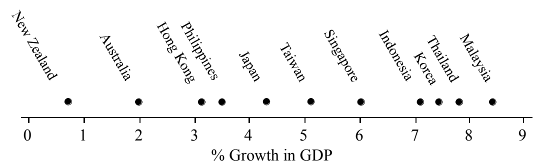


Figure 2.3.5 Forecast of percent growth in GDP for 1990 for some South-East Asian and Pacific countries.

Example of a stem-and-leaf plot

Units: 7 | 2 =72

Stem	Leafs
1	7
2	7 9
3	0 7 7 8 9
4	2 2 3 3 4 5 8 8 9
5	0 0 0 1 1 2 2 3 4 4 7 8 9 9
6	0 0 0 2 3 3 3 3 6 7 7 8
7	2 2

Values 52, 54 and their frequencies

Stem-plot of the 45 obs's of the Ejection variable in the Heart Attack data table.

Traffic death-rates data

TABLE 2.3.1 Traffic Death-Rates (per 100,000 Population) for 30 Countries

17.4 Australia	20.1 Austria	19.9 Belgium	12.5 Bulgaria	15.8 Canada
10.1 Czechoslovakia	13.0 Denmark	11.6 Finland	20.0 France	12.0 E. Germany
13.1 W. Germany	21.1 Greece	5.4 Hong Kong	17.1 Hungary	15.3 Ireland
10.3 Israel	10.4 Japan	26.8 Kuwait	11.3 Netherlands	20.1 New Zealand
10.5 Norway	14.6 Poland	25.6 Portugal	12.6 Singapore	9.8 Sweden
15.7 Switzerland	18.6 United States	12.1 N. Ireland	12.0 Scotland	10.1England & Wales

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)

Source: Hutchinson [1987, page 3].

Units: 17 | 4 = 17.4 deaths per 100,000

5	4
6	
7	
8	
9	8
10	1 1 3 4 5
11	3 6
12	0 0 1 5 6
13	0 1
14	6
15	3 7 8
16	
17	1 4
18	6
19	9
20	0 1 1
21	1
22	
23	
24	
25	6
26	8

(a)

Round-off

Collapse to

12 stems

Units: 1 | 7 = 17 deaths per 100,000

0	5
0	
0	
1	0 0 0 0 0 1 1
1	2 2 2 2 3 3 3
1	5 5
1	6 6 7 7
1	9
2	0 0 0 0 1
2	
2	
2	6 7

(b)

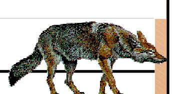
FIGURE 2.3.7 Two stem-and-leaf plots for the traffic deaths dat

TABLE 2.3.2 Coyote Lengths Data (cm)

Females											
93.0	97.0	92.0	101.6	93.0	84.5	102.5	97.8	91.0	98.0	93.5	91.7
90.2	91.5	80.0	86.4	91.4	83.5	88.0	71.0	81.3	88.5	86.5	90.0
84.0	89.5	84.0	85.0	87.0	88.0	86.5	96.0	87.0	93.5	93.5	90.0
85.0	97.0	86.0	73.7								
Males											
97.0	95.0	96.0	91.0	95.0	84.5	88.0	96.0	96.0	87.0	95.0	100.0
101.0	96.0	93.0	92.5	95.0	98.5	88.0	81.3	91.4	88.9	86.4	101.6
83.8	104.1	88.9	92.0	91.0	90.0	85.0	93.5	78.0	100.5	103.0	91.0
105.0	86.0	95.5	86.5	90.5	80.0	80.0					

Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

TABLE 2.3.3 Frequency Table for Female Coyote Lengths

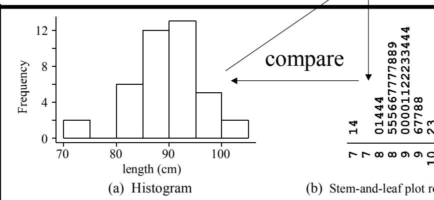


Body length

Class Interval	Tally	Frequency	Stem	Stem-and-leaf plot
70-75⁻		2	7	1 4
75-80⁻		0	7	
80-85⁻		6	8	0 1 4 4 4
85-90⁻		12	8	5 5 5 6 6 7 7 7 7 8 8 9
90-95⁻		13	9	0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100⁻		5	9	6 7 7 8 8
100-105⁻		2	10	2 3
Total		40		

TABLE 2.3.3 Frequency Table for Female Coyote Lengths

Class Interval	Tally	Frequency	Stem	Stem-and-leaf plot
70-75⁻		2	7	1 4
75-80⁻		0	7	
80-85⁻		6	8	0 1 4 4 4
85-90⁻		12	8	5 5 5 6 6 7 7 7 7 8 8 9
90-95⁻		13	9	0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100⁻		5	9	6 7 7 8 8
100-105⁻		2	10	2 3
Total		40		



(a) Histogram (b) Stem-and-leaf plot rotated

Figure 2.3.8 Histogram of the female coyote-lengths data.

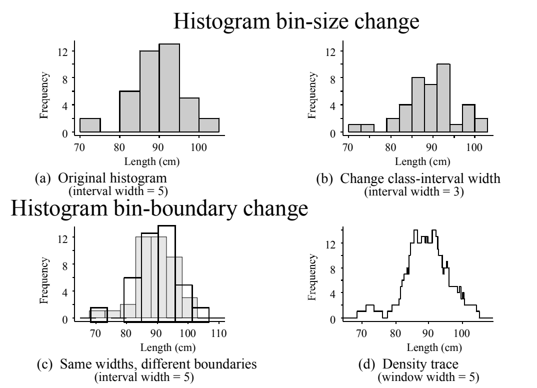
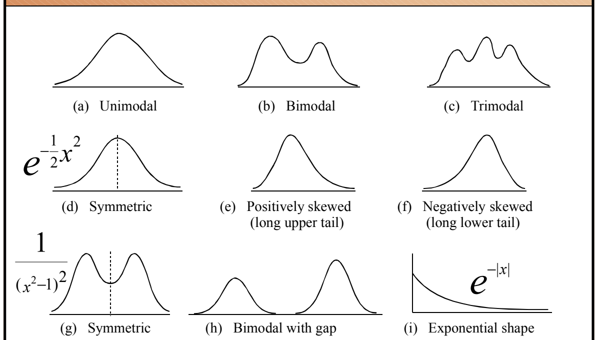


Figure 2.3.9 Histograms and density trace of female coyote-lengths data.

Questions …

- What advantages does a stem-and-leaf plot have over a histogram? (S&L Plots return info on individual values, quick to produce by hand, provide data sorting mechanisms. But, histograms are more attractive and more understandable).
- The shape of a histogram can be quite drastically altered by choosing different class-interval boundaries. What type of plot does not have this problem? (density trace) What other factor affects the shape of a histogram? (bin-size)
- What was another reason given for plotting data on a variable, apart from interest in how the data on that variable behaves? (shows features, cluster/gaps, outliers; as well as trends)

Interpreting Stem-plots and Histograms



Interpreting Stem-plots and Histograms

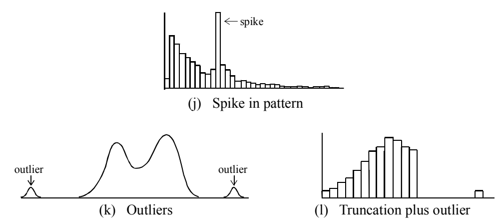
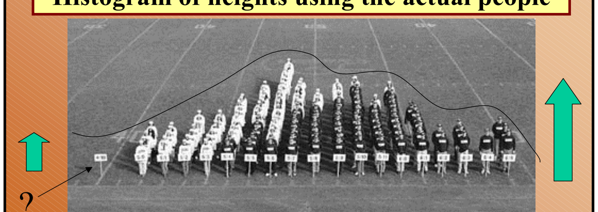


Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

Fascinations with histograms – Histogram of heights using the actual people



Subjects are university genetics students, females in white and males in dark tops.

Skewness & Kurtosis

- What do we mean by symmetry and positive and negative skewness? Kurtosis? Properties?!?

$$\text{Skewness} = \frac{\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^3}{(N-1)SD^3}; \quad \text{Kurtosis} = \frac{\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^4}{(N-1)SD^4}$$

- Skewness in linearly invariant Sk(aX+b)=Sk(X)
- Skewness is a measure of unsymmetry
- Kurtosis is a measure of flatness
- Both are use to quantify departures from StdNormal
- Skewness(StdNorm)=0; Kurtosis(StdNorm)=3

Descriptive statistics from computer programs like STATA

STATA Output

Standard deviation

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
age	45	50.133	51.000	50.366	6.092	0.908

Variable	Minimum	Maximum	Q1	Q3
age	36.000	59.000	46.500	56.000

Lower quartile *Upper quartile*

Descriptive statistics …

- The sample mean is denoted by $\overline{x}$.

$$\text{The } \textbf{\textit{sample mean}} = \frac{\textbf{\textit{Sum}} \text{ of the observations}}{\textbf{\textit{Number}} \text{ of observations}}$$

The sample mean is where the dot plot balances

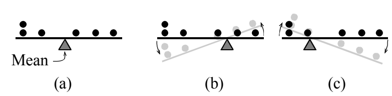


Figure 2.4.1 Mechanical construction representing a dot plot: (a) shows a balanced rod while (b) and (c) show unbalanced rods.

The sample median

For n observations, {x1, x2, x3, …, xn}. Suppose we order the observations min-to-max to get

{x(1), x(2), x(3), …, x(n)}.

Then the ***sample median*** is the [(n+1)/2]-st largest Observation $x_{\left(\frac{n+1}{2}\right)}$

If $\frac{n+1}{2}$ is not a whole number, the median is the average of the two observations on either side.

Effect of outliers on the mean and median

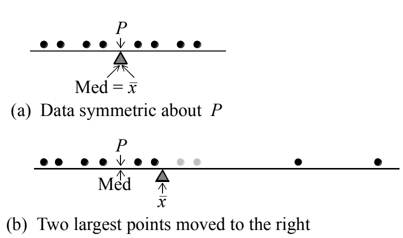
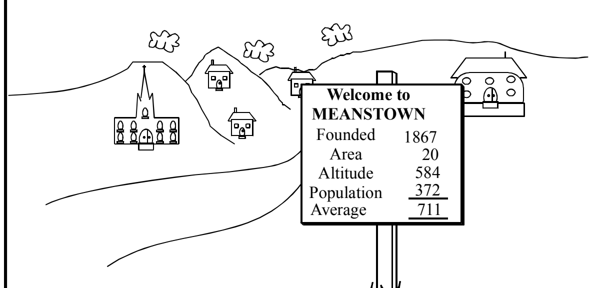


Figure 2.4.2 The mean and the median.
[Grey disks in (b) are the ``ghosts" of the points that were moved.]

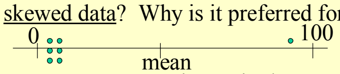
Beware of inappropriate averaging



Suggested by a 1977 cartoon in *The New Yorker* magazine by Dana Fradon.

Questions …

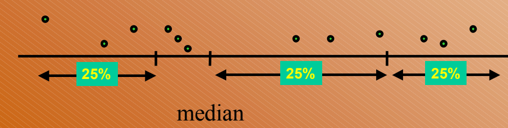
- How is the sample mean related to the dot plot?
- If the index $(n+1)/2$ is not a whole number (e.g., 23.5), how do we obtain the sample median?
- Why is the sample median usually preferred to the sample mean for skewed data? Why is it preferred for "dirty" data?



- Under what circumstances may quoting a single center (be it mean or median) not make sense?(multi-modal)
- What can we say about the sample mean of a qualitative variable? (meaningless)

Quartiles

The first quartile (Q_1) is the median of all the observations whose *position* is strictly below the *position* of the median, and the third quartile (Q_3) is the median of those above.



Mode, Coefficient-of-Variation

•Mode: the most frequently occurring number in a discrete data sample.

•CV: Coefficient of variation = SD/Mean

Five number summary

The five-number summery = (Min, Q_1, Med, Q_3, Max)

Inter-quartile Range

$$IQR = Q_3 - Q_1$$

Box plot compared to dot plot

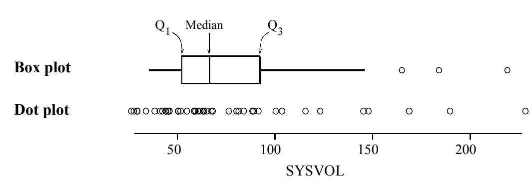


Figure 2.4.3 Box plot for SYSVOL.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Construction of a box plot

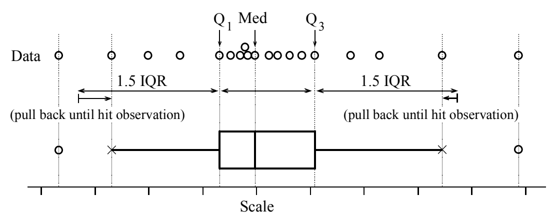


Figure 2.4.4 Construction of a box plot.

Frequency Table

TABLE 2.5.1 Word Lengths for the First 100 Words on a Randomly Chosen Page

3	2	2	4	4	4	3	9	9	3	6	2	3	2	3	4	6	5	3	4
2	3	4	5	2	9	5	8	3	2	4	5	2	4	1	4	2	5	2	5
3	6	9	6	3	2	3	4	4	4	2	2	4	2	3	7	4	2	6	4
2	5	9	2	3	7	11	2	3	6	4	4	7	6	6	10	4	3	5	7
7	7	5	10	3	2	3	9	4	5	5	4	4	3	5	2	5	2	4	2

Frequency Table

Value u_j	1	2	3	4	5	6	7	8	9	10	11
Frequency f_j	1	22	18	22	13	8	6	1	6	2	1

Mean from a frequency table

$$\bar{x} = \frac{1}{n}\text{Sum of (value} \times \text{frequency of occurrence)} =$$

$$\frac{1}{n}\text{(Sum of all observations)}$$

TABLE 2.5.2
Frequency Table for the Occurrence of Fish Species in Ocean Strata

No. of strata in which species occur (u_j)	Frequency (No. of species) (f_j)	Percentage of species ($\frac{f_j}{n} \times 100$)	Cumulative Percentage
1	117	35.5	35.5
2	61	18.5	53.9
3	37	11.2	65.2
4	24	7.3	72.4
5	23	7.0	79.4
6	12	3.6	83.0
7	14	4.2	87.3
8	10	3.0	90.3
9	9	2.7	93.0
10+	23	7.0	100.0
	n = 330	100	

Source: Haedrich and Merrett [1988]

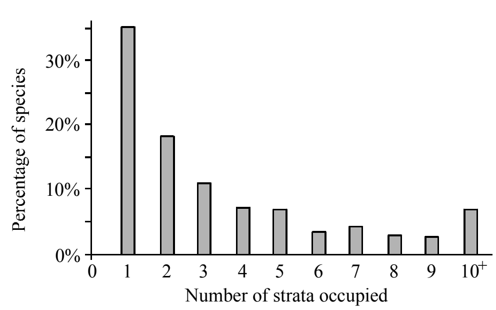


Figure 2.5.1 Bar graph for species data.

Sampling Distributions

- Parameters and Estimates
- Sampling distributions of the sample mean
- Central Limit Theorem (CLT)
- Estimates that are approximately Normal
- Standard errors of differences
- Student's *t*-distribution

Parameters and estimates

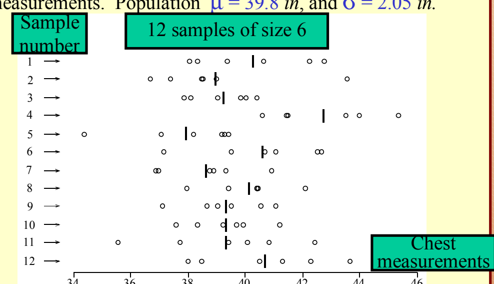
- A ***parameter*** is a numerical characteristic of a population or distribution
- An ***estimate*** is a quantity calculated from the data to <u>approximate</u> an **unknown** parameter
- Notation
  - Capital letters refer to random variables
  - Small letters refer to observed values

Questions

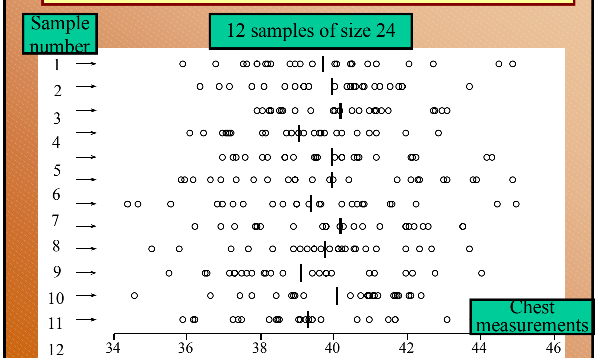
- What are two ways in which random observations arise and give examples. (random sampling from finite population – randomized scientific experiment; random process producing data.)
- What is a parameter? Give two examples of parameters. (characteristic of the data – mean, 1st quartile, std.dev.)
- What is an estimate? How would you estimate the parameters you described in the previous question?
- What is the distinction between an estimate (p^ value calculated form obs'd data to approx. a parameter) and an estimator (P^ abstraction the the properties of the ransom process and the sample that produced the estimate) ? Why is this distinction necessary? (effects of sampling variation in P^)

The sample mean has a sampling distribution

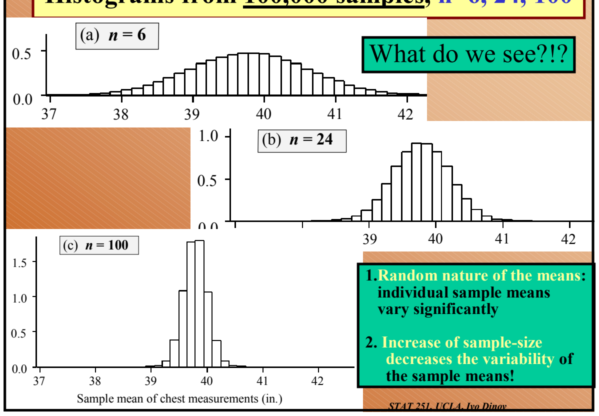
Sampling batches of Scottish soldiers and taking chest measurements. Population $\mu = 39.8$ *in*, and $\sigma = 2.05$ *in*.



Twelve samples of size 24



Histograms from <u>100,000 samples</u>, n=6, 24, 100



1. **Random nature of the means: individual sample means vary significantly**
2. **Increase of sample-size decreases the variability of the sample means!**

Mean and SD of the sampling distribution of X_bar

[Sampling distributions -probability distributions of statistics]

E(*sample mean*) = Population *mean*

$$\mathrm{SD}(\textit{sample mean}) = \frac{\text{Population } SD}{\sqrt{\text{Sample size}}}$$

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X) = \mu, \quad \mathrm{SD}(\bar{X}) = \frac{\mathrm{SD}(X)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

Review

- We use both $\bar{x}$ and $\bar{X}$ to refer to a sample mean. For what purposes do we use the *former* and for what purposes do we use the *latter*?
- What is meant by "the sampling distribution of $\bar{X}$"?

  (sampling variation – the observed variability in the process of taking random samples; sampling distribution – the real probability distribution of the random sampling process)
- How is the population mean of the sample average $\bar{X}$ related to the population mean of individual observations? (E($\bar{X}$) = Population mean)

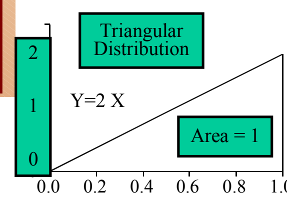
Review

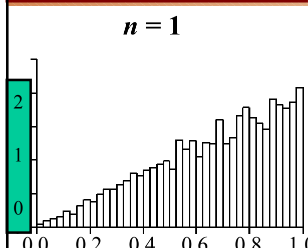
- How is the population standard deviation of $\bar{X}$ related to the population standard deviation of individual observations? (SD($\bar{X}$) = (Population SD)/sqrt(sample_size))
- What happens to the sampling distribution of $\bar{X}$ if the sample size is increased? (variability decreases)
- What does it mean when $\bar{x}$ is said to be an "unbiased estimate" of μ ? (E($\bar{x}$) = μ. Are Y^= ¼ *Sum*, or Z^ = ¾ *Sum* unbiased?)
- If you sample from a Normal distribution, what can you say about the distribution of $\bar{X}$? (Also Normal)

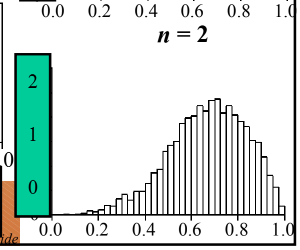
Review

- Increasing the precision of $\bar{X}$ as an estimator of μ is equivalent to doing what to SD($\bar{X}$)? (decreasing)
- For the sample mean calculated from a random sample, SD($\bar{X}$) = $\frac{\sigma}{\sqrt{n}}$. This implies that the variability from sample to sample in the *sample-means* is given by the variability of the individual observations divided by the square root of the sample-size. In a way, averaging decreases variability.

Central Limit Effect – Histograms of sample means



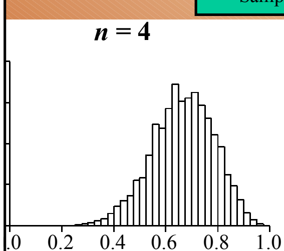


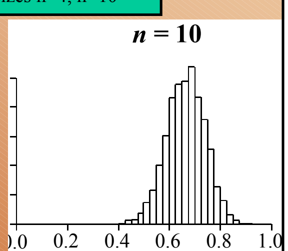


Sample means from sample size n=1, n=2, 500 samples

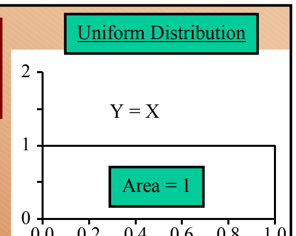
Central Limit Effect -- Histograms of sample means

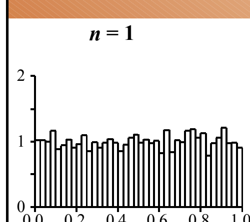
Triangular Distribution
Sample sizes n=4, n=10

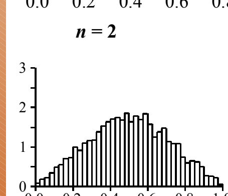




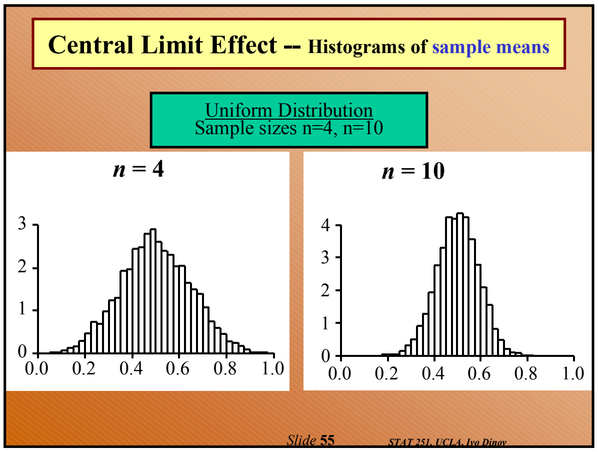
Central Limit Effect – Histograms of sample means



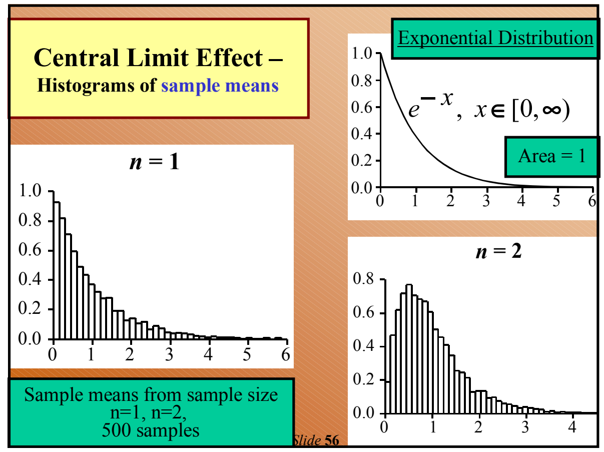




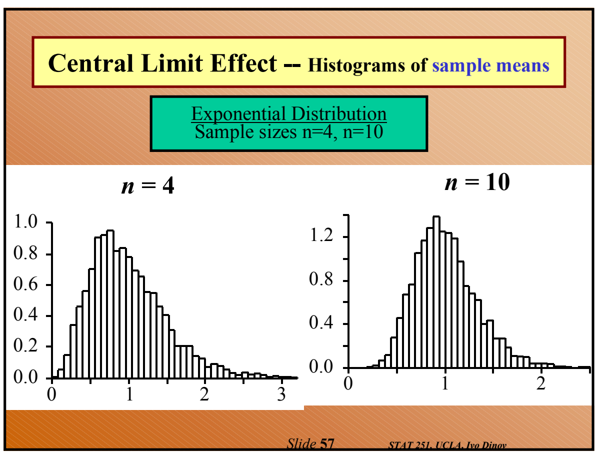
Sample means from sample size n=1, n=2, 500 samples



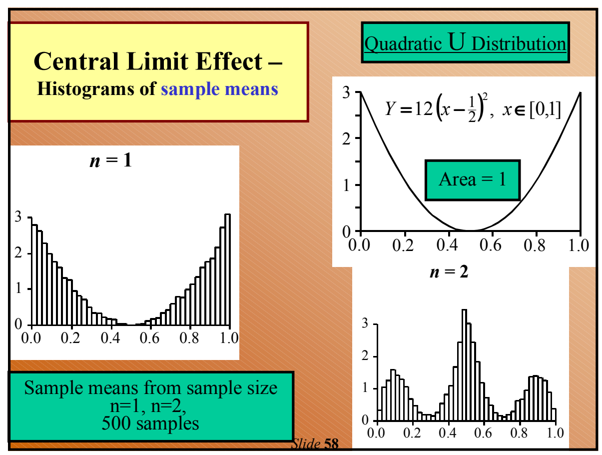
Central Limit Effect -- Histograms of sample means
Uniform Distribution
Sample sizes n=4, n=10
n = 4
n = 10
Slide 55
STAT 251, UCLA, Ivo Dinov



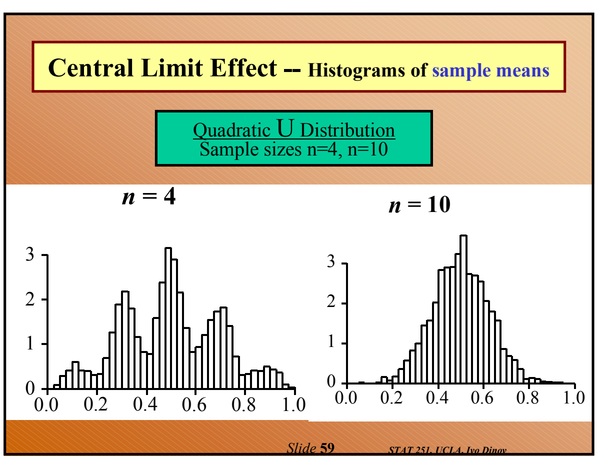
Central Limit Effect –
Histograms of sample means
Exponential Distribution
$e^{-x},\ x \in [0,\infty)$
Area = 1
n = 1
n = 2
Sample means from sample size
n=1, n=2,
500 samples
Slide 56



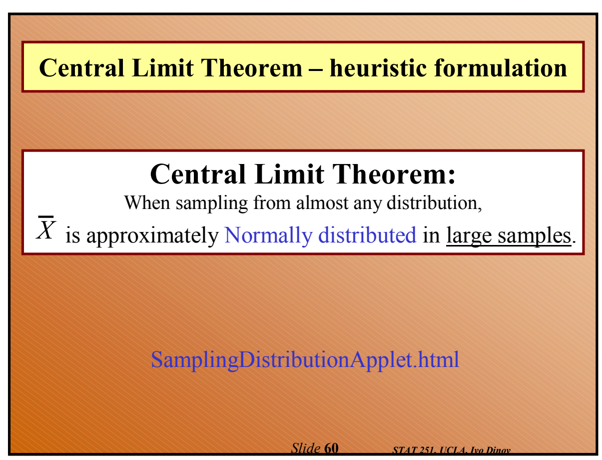
Central Limit Effect -- Histograms of sample means
Exponential Distribution
Sample sizes n=4, n=10
n = 4
n = 10
Slide 57
STAT 251, UCLA, Ivo Dinov



Central Limit Effect –
Histograms of sample means
Quadratic U Distribution
$Y = 12\left(x-\frac{1}{2}\right)^2,\ x \in [0,1]$
Area = 1
n = 1
n = 2
Sample means from sample size
n=1, n=2,
500 samples
Slide 58



Central Limit Effect -- Histograms of sample means
Quadratic U Distribution
Sample sizes n=4, n=10
n = 4
n = 10
Slide 59
STAT 251, UCLA, Ivo Dinov



Central Limit Theorem – heuristic formulation
Central Limit Theorem:
When sampling from almost any distribution,
$\overline{X}$ is approximately Normally distributed in large samples.
SamplingDistributionApplet.html
Slide 60
STAT 251, UCLA, Ivo Dinov

Central Limit Theorem – theoretical formulation

Let $\{X_1, X_2, \ldots, X_k, \ldots\}$ be a sequence of independent observations from one specific random process. Let and $E(X) = \mu$ and $SD(X) = \sigma$ and both are finite ($0 < \sigma < \infty$; $|\mu| < \infty$). If $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, sample-avg,

Then $\overline{X}$ has a distribution which approaches N(μ, σ^2/n), as $n \to \infty$.

Review

- What does the central limit theorem say? Why is it useful? (If the sample sizes are large, the mean in Normally distributed, as a RV)
- In what way might you expect the central limit effect to differ between samples from a symmetric distribution and samples from a very skewed distribution? (Larger samples for non-symmetric distributions to see CLT effects)
- What other important factor, apart from skewness, slows down the action of the central limit effect?

(Heavyness in the tails of the original distribution.)

Review

- When you have data from a moderate to small sample and want to use a normal approximation to the distribution of $\overline{X}$ in a calculation, what would you want to do before having any faith in the results? (30 or more for the sample-size, depending on the skewness of the distribution of *X*. Plot the data - non-symmetry and heavyness in the tails slows down the CLT effects).
- Take-home message: CLT is an application of statistics of paramount importance. Often, we are not sure of the distribution of an observable process. However, the CLT gives us a theoretical description of the distribution of the sample means as the sample-size increases (N(μ, σ^2/n)).

The standard error of the mean – remember …

- For the sample mean calculated from a random sample, SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$. This implies that the variability from sample to sample in the *sample-means* is given by the variability of the individual observations divided by the square root of the sample-size. In a way, averaging decreases variability.
- Recall that for *known* SD(X)=σ, we can express the SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$. How about if SD(X) is *unknown*?!?

The standard error of the mean

The standard error of the sample mean is an estimate of the *SD* of the sample mean

- i.e. a measure of the precision of the sample mean as an estimate of the population mean
- given by $SE(\overline{x}) = \frac{\text{Sample standard deviation}}{\sqrt{\text{Sample size}}}$

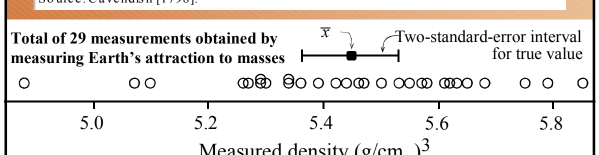
$$SE(\overline{x}) = \frac{s_x}{\sqrt{n}}$$

- Note similarity with
- SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$.

Cavendish's 1798 data on mean density of the Earth, g/cm^3, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].



Newton's law of gravitation: F = G m_1 m_2 /r^2, the attraction force F is the ratio of the product (Gravitational const, mass of body1, mass body2) and the distance between them, r. Goal is to estimate G!

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].

Sample mean $\bar{x} = 5.447931 \ g/cm^3$

and sample SD = $s_X = 0.2209457 \ g/cm^3$

Then the standard error for these data is:

$$SE(\bar{X}) = \frac{s_x}{\sqrt{n}} = \frac{0.2209457}{\sqrt{29}} = 0.04102858$$

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].

Safely can assume the true mean density of the Earth is within 2 SE's of the sample mean!

$$\bar{x} \pm 2 \times SE(\bar{x}) = 5.447931 \pm 2 \times 0.04102858 \, g/cm^3$$

Review

- Why is the standard deviation of $\bar{X}$, SD($\bar{X}$), not a useful measure of the precision of $\bar{X}$ as an estimator in practical applications? (SD($\bar{X}$) = $\frac{\sigma}{\sqrt{n}}$ and σ is unknown most time!)
- What measure of precision do we use in practice? (SE)
- How is SE($\bar{x}$) related to SD($\bar{X}$)?
- When we use the formula SE($\bar{x}$) = $s_X / \sqrt{n}$, what is s_X and how do you obtain it? (Sample SD(X))

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

Review

- What can we say about the true value of μ and the interval $\bar{x} \pm 2$ SE($\bar{x}$) ? (95% sure)
- Increasing the precision of $\bar{x}$ as an estimate of μ is equivalent to doing what to se($\bar{x}$)? (decreasing)

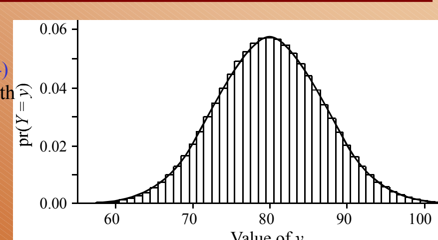
Sampling distribution of the sample proportion

The sample proportion $\hat{p}$ estimates the population proportion p.

Suppose, we poll college athletes to see what percentage are using performance enhancing drugs. If 25% admit to using such drugs (in a single poll) can we trust the results? What is the variability of this proportion measure (over multiple surveys)? Could Football, Water Polo, Skiing and Chess players have the same drug usage rates?

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p)



$\mu_Y = E(Y) = np$

$\sigma_Y = SD(Y) = \sqrt{np(1-p)}$

For large samples, the distribution of $\hat{P}$ is approximately Normal with mean $= p$ and standard deviation $= \sqrt{\frac{p(1-p)}{n}}$

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p). Y = # Heads in n-trials. Hence, the proportion of Heads is: Z=Y/n.

$\mu_Y = E(Y) = np$ $\qquad \mu_Z = E(Z) = \frac{1}{n}E(Y) = p$

$\sigma_Y = SD(Y) = \sqrt{np(1-p)}$ $\qquad \sigma_Z = SD(Z) = \frac{1}{n}SD(Y) = \sqrt{\frac{p(1-p)}{n}}$

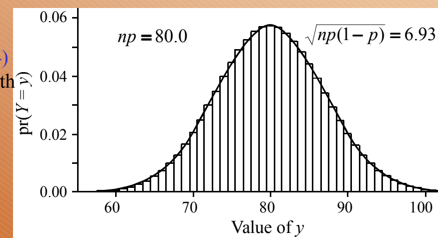
This gives us bounds on the variability of the sample proportion:

$$\mu_Z \pm 2SE(Z) = p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

What is the variability of this proportion measure over multiple surveys?

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p)



The sample proportion Y/n can be approximated by normal distribution, by CLT, and this explains the tight fit between the observed histogram and a N(pn, $\sqrt{np(1-p)}$)

Standard error of the sample proportion

Standard error of the sample proportion:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Review

- We use both $\hat{p}$ and $\hat{P}$ to describe a sample proportion. For what purposes do we use the former and for what purposes do we use the latter? (observed values vs. RV)
- What two models were discussed in connection with investigating the distribution of $\hat{p}$? What assumptions are made by each model? (Number of units having a property from a large population Y~ Bin(n,p), when sample <10% of popul.; Y/n~Normal(m,s), since it's the avg. of all Head(1) and Tail(0) observations, when n-large).
- What is the standard deviation of a sample proportion obtained from a binomial experiment?

$$SD(Y/n) \equiv \sqrt{\frac{p(1-p)}{n}}$$

Review

- Why is the standard deviation of $\hat{P}$ not useful in practice as a measure of the precision of the estimate? $SD(\hat{P}) \equiv \sqrt{\frac{p(1-p)}{n}}$, in terms of p = unknown!
- How did we obtain a useful measure of precision, and what is it called? (SE($\hat{p}$))
- What can we say about the true value of p and the interval $\hat{p} \pm 2$ SE($\hat{p}$)? (Safe bet!)
- Under what conditions is the formula

  SE($\hat{p}$) $= \sqrt{\hat{p}(1-\hat{p})/n}$ applicable? (Large samples)

Review

- In the TV show Annual People's Choice Awards, awards are given in many categories (including favorite TV comedy show, and favorite TV drama) and are chosen using a Gallup poll of 5,000 Americans (US population approx. 260 million).
- At the time the 1988 Awards were screened in NZ, an NZ Listener journalist did "a bit of a survey" and came up with a list of awards for NZ (population 3.2 million).
- Her list differed somewhat from the U.S. list. She said, *"it may be worth noting that in both cases approximately 0.002 percent of each country's populations were surveyed."* The reporter inferred that because of this fact, her survey was just as reliable as the Gallup poll. Do you agree? Justify your answer. (only 62 people surveyed, but that's okay. Possible bad design (not a random sample)?)

Review

- Are public opinion polls involving face-to-face interviews typically simple random samples? (No! Often there are elements of quota sampling in public opinion polls. Also, most of the time, samples are taken at random from clusters, e.g., townships, counties, which doesn't always mean random sampling. Recall, however, that the size of the sample doesn't really matter, as long as it's random, since sample size less than 10% of population implies Normal approximation to Binomial is valid.)
- What approximate measure of error is commonly quoted with poll results in the media? What poll percentages does this level of error apply to?

  ($\hat{p} \pm 2*SE(\hat{p})$, 95%, from the Normal approximation)

Review

- A 1997 questionnaire investigating the opinions of computer hackers was available on the internet for 2 months and attracted 101 responses, e.g. 82% said that stricter criminal laws would have no effect on their activities. Why would you have no faith that a 2 std-error interval would cover the true proportion?

  (sampling errors present (self-selection), which are a lot larger than non-sampling statistical random errors).

Bias and Precision

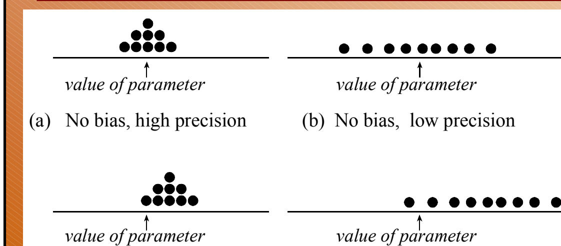
- The bias in an estimator is the distance between between the *center* of the sampling distribution of the estimator and the true value of the parameter being estimated. In math terms, bias = $E(\hat{\Theta}) - \theta$, where theta $\hat{\Theta}$ is the estimator, as a RV, of the true (unknown) parameter θ .
- Example, Why is the sample mean an **unbiased** estimate for the population mean? How about ¾ of the sample mean?

$$E(\hat{\Theta}) - \mu = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) - \mu = 0$$

$$E(\hat{\Theta}) - \mu = E\left(\frac{3}{4}\frac{1}{n}\sum_{k=1}^{n} X_k\right) - \mu = \frac{3}{4}\mu - \mu = \frac{\mu}{4} \neq 0, \quad \text{in general.}$$

Bias and Precision

- The precision of an estimator is a measure of how variable is the estimator in repeated sampling.



(a) No bias, high precision (b) No bias, low precision

(c) Biased, high precision (d) Biased, low precision

Standard error of an estimate

The ***standard error*** of any estimate $\hat{\theta}$ [denoted se($\hat{\theta}$)]

- estimates the variability of $\hat{\theta}$ values in repeated sampling and
- is a measure of the ***precision*** of $\hat{\theta}$.

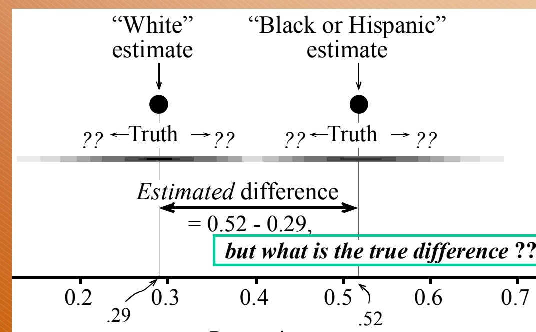
Review

- What is meant by the terms parameter and estimate.
- Is an estimator a RV?
- What is statistical inference? (process of making conclusions or making useful statements about unknown distribution parameters based on observed data.)
- What are bias and precision?
- What is meant when an estimate of an unknown parameter is described as unbiased?

Review

- What is the standard error of an estimate, and what do we use it for? (measure of precision)
- Given that an estimator of a parameter is approximately normally distributed, where can we expect the true value of the parameter to lie? (within 2SE away)
- If each of 1000 researchers independently conducted a study to estimate a parameter θ, how many researchers would you expect to catch the true value of θ in their 2-standard-error interval? (10*95=950)

Estimating a difference – proportions of people who believe police use racial profiling



Is there racial profiling or are there confounding explanatory effects?!?

- The book by Best (*Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists*, Joel Best) shows how we can test for racial bias in police arrests. Suppose we find that among 100 white and 100 black youths, 10 and 17, respectively, have experienced arrest. This may **look plainly discriminatory**. But suppose we then find that of the 80 middle-class white youths 4 have been arrested, and of the 50 middle-class black youths 2 arrested, whereas the corresponding numbers of lower-class white and black youths arrested are, respectively, 6 of 20 and 15 of 50. These arrest rates correspond to 5 per 100 for white and 4 per 100 for black middle-class youths, and 30 per 100 for both white and black lower-class youths. Now, better analyzed, the data suggest **effects of social class, not race as such**.

Standard error of a difference

Standard error for a difference between independent estimates:

$$SE(Est_1 - Est_2) = \sqrt{SE(Est_1)^2 + SE(Est_2)^2}$$

or

$$SE(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{SE(\hat{\theta}_1)^2 + SE(\hat{\theta}_2)^2}$$

What happens if there's an association between them?

Student's *t*-distribution

- For random samples from a Normal distribution,

$$T = \frac{(\overline{X} - \mu)}{SE(\overline{X})}$$

Recall that for samples from N(μ , σ)

$$Z = \frac{(\overline{X} - \mu)}{SD(\overline{X})} = \frac{(\overline{X} - \mu)}{\sigma / \sqrt{n}} \sim N(0,1)$$

is exactly distributed as Student($df = n - 1$) ← Approx/Exact Distributions

- but methods we shall base upon this distribution for T work well even for small samples sampled from distributions which are quite non-Normal.
- df is number of observations –1, degrees of freedom.

Density curves for Student's *t*

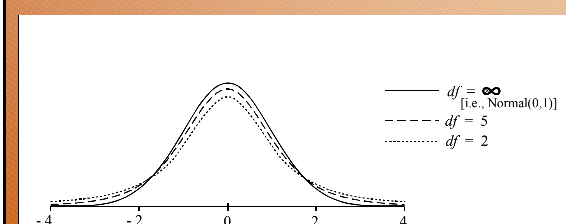


Figure 7.6.1 Student(df) density curves for various *df*.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Notation

- By $t_{df}(prob)$, we mean the number t such that when $T \sim$ Student(df), P($T \geq t_{df}$) = $prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.

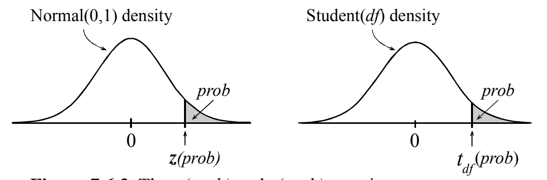


Figure 7.6.2 The $z(prob)$ and $t(prob)$ notations.

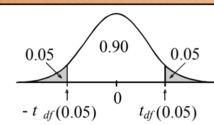
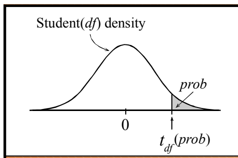


Figure 7.6.3 The central 90% of the Student(df) distribution.

Reading Student's *t* table



Desired *upper-tail prob*

TABLE 7.6.1 Extracts from the Student's t-Distribution Table

prob

df	.20	.15	.10	.05	.025	.01	.005	.001	.0005	.0001
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
...	...	...	...	...	...	...	...	...	...	...
10	0.879	1.093	1.372	1.812	$\boxed{2.228}$	2.764	3.169	4.144	4.587	5.694
...	...	...	...	...	...	...	...	...	...	...
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
...	...	...	...	...	...	...	...	...	...	...
∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

Desired *df*

t-value

Review

- Qualitatively, how does the Student (df) distribution differ from the standard Normal(0,1) distribution? What effect does increasing the value of df have on the shape of the distribution? (σ is replaced by SE)
- What is the relationship between the Student ($df = \infty$) distribution and the Normal(0,1) distribution? (Approximates N(0,1) as n→increases)

Review

- Why is T, the number of standard errors separating $\overline{X}$ and μ , a more variable quantity than Z, the number of standard deviations separating $\overline{X}$ and μ ? (Since an additional source of variability is introduced in T, SE, not available in Z. E.g., P(-2<=T<=2)=0.9144 < 0.954=P(-2<=Z<2), hence tails of T are wider. To get 95% confidence for T we need to go out to +/-2.365).
- For large samples the true value of μ lies inside the interval $\bar{x} \pm 2$ se($\bar{x}$) for a little more than 95% of all samples taken. For small samples from a normal distribution, <u>is the proportion of samples for which the true value of μ lies within the 2-standard-error interval smaller or bigger than 95%?</u> Why?(Smaller – wider tail.)

Review

- For a small Normal sample, if you want an interval to contain the true value of μ for 95 % of samples taken, should you take more or fewer than two-standard errors on either side of $\bar{x}$? (more)
- Under what circumstances does mathematical theory show that the distribution of T=($\bar{X}$- μ)/SE($\overline{X}$) is exactly Student (df=n-1)? (Normal samples)
- Why would methods derived from the theory be of little practical use if they stopped working whenever the data was not normally distributed? (In practice, we're never sure of Normality of our sampling distribution).

Sampling Distributions

- For random quantities, we use a capital letter for the random variable, and a small letter for an observed value, for example, X and x, $\overline{X}$ and $\overline{x}$, $\hat{P}$ and $\hat{p}$, $\hat{\Theta}$ and $\hat{\theta}$.
- In estimation, the random variables (capital letters) are used when we want to think about the effects of sampling variation, that is, about how the random process of taking a sample and calculating an estimate behaves.

Sampling distribution of $\overline{X}$

Sample mean, $\overline{X}$:

For a random sample of size n from a distribution for which $E(X) = \mu$ and $sd(X) = \sigma$, the sample mean $\overline{X}$ has :

- $E(\overline{X}) = E(X) = \mu$, $SD(\overline{X}) = \frac{SD(X)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$
- If we are sampling from a Normal distribution, then $\overline{X} \sim$ Normal. (**exactly**)
- *Central Limit Theorem*: For almost any distribution, $\overline{X}$ is **approximately** Normally distributed in large samples.

Sampling distribution of the sample proportion

- **Sample proportion, $\hat{P}$:** For a random sample of size n from a population in which a proportion p have a characteristic of interest, we have the following results about the sample proportion with that characteristic:
  - $\mu_{\hat{p}} = E(\hat{P}) = p$ $\sigma_{\hat{p}} = sd(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$
  - $\hat{P}$ is approximately Normally distributed for large n

(e.g., $np(1-p) \geq 10$, though a more accurate rule is given in the next chapter)

Parameters and estimates

- A ***parameter*** is a numerical characteristic of a population or distribution
- An ***estimate*** is a known quantity calculated from the data to approximate an unknown parameter
  - For general discussions about parameters and estimates, we talk in terms of $\hat{\theta}$ being an estimate of a parameter θ
  - The ***bias*** in an estimator is the difference between $E(\hat{\Theta})$ and θ
  - $\hat{\theta}$ is an ***unbiased estimate*** of θ if $E(\hat{\Theta}) = \theta$.

Precision

- The ***precision*** of an estimate refers to its variability in repeated sampling
- One <u>estimate is less precise than another</u> if it has more *variability*.

Standard error

- **The standard error**, SE($\hat{\theta}$), for an estimate $\hat{\theta}$ is:
  - an estimate of the std dev. of the sampling distribution
  - a measure of the precision of $\hat{\theta}$ as an estimate of θ
- **For a mean**
  - The sample mean $\overline{x}$ is an unbiased estimate of the population mean μ
  - $SE(\overline{x}) = \frac{s_X}{\sqrt{n}}$

Standard errors cont.

- **Proportions**
  - The sample proportion $\hat{p}$ is an unbiased estimate of the population proportion p
  - $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Standard error of a difference:** For independent estimates,

$$\mathrm{se}(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\mathrm{se}(\hat{\theta}_1)^2 + \mathrm{se}(\hat{\theta}_2)^2}$$

TABLE 7.7.1 Some Parameters and Their Estimates

	Population(s) or Distributions(s) → Parameters	Sample data → Estimates	→ Measure of precision
Mean	m	$\bar{x}$	se $(\bar{x})$
Proportion	p	$\hat{p}$	se $(\hat{p})$
Difference in means	μ_1-μ_2	$\bar{x}_1 - \bar{x}_2$	se $(\bar{x}_1 - \bar{x}_2)$
Difference in proportions	p_1-p_2	$\hat{p}_1 - \hat{p}_2$	se $(\hat{p}_1 - \hat{p}_2)$
General case	θ	$\hat{\theta}$	se $(\hat{\theta})$

Student's *t*-distribution ….

- Is bell shaped and centered at zero like the Normal(0,1), but
- More variable (larger spread and fatter tails).
- As *df* becomes larger, the Student(*df*) distribution becomes more and more like the Normal(0,1) distribution.
- Student$(df = \infty)$ and Normal(0,1) are two ways of describing the same distribution.

Student's *t*-distribution cont.

- For random samples from a Normal distribution,

$$T = (\bar{X} - \mu) / SE(\bar{X})$$

  is exactly distributed as Student($df = n$ - 1), but methods we shall base upon this distribution for T work well even for small samples sampled from distributions which are quite non-Normal.
- By $t_{df}(prob)$, we mean the number t such that when $T \sim$ Student(df), pr($T \geq t$) = $prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.

CLT Example – CI shrinks by half by quadrupling the sample size!

- If I ask 30 of you the question "Is 3 credit hour a reasonable load for Stat251?", and say, 15 (50%) said ***no***. Should we change the format of the class?
- Not really – the 2SE interval is about **[0.32 ; 0.68]**. So, we have little concrete evidence of the proportion of students who think we need a change in Stat 251 format,

$$\hat{p} \pm 2 \times \mathrm{SE}(\hat{p}) = 0.5 \pm 2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.5 \pm -0.18$$

- If I ask all 300 Stat 251 students and 150 say ***no*** (still 50%), then 2SE interval around 50% is: **[0.44 ; 0.56]**.
- **So, large sample is much more useful and this is due to CLT effects, without which, we have no clue how useful our estimate actually is …**