

UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor:** Ivo Dinov,
Asst. Prof. In Statistics and Neurology
- **Teaching Assistants:**
Ming Zheng, Annie Che
UCLA Statistics

University of California, Los Angeles, Winter 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13, UCLA, Ivo Dinov Slide 1

Statistics Online Compute Resources

- <http://socr.stat.ucla.edu/SOCR.html>
- Interactive Normal Curve
- Online Calculators for Binomial, Normal, Chi-Square, F and T distributions
- Galton's Board or *Quincunx*
- *BinomialCoinExperiment, DieExperiment*

STAT 13, UCLA, Ivo Dinov Slide 2

Chapter 7: 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

STAT 13, UCLA, Ivo Dinov Slide 3

Parameters and estimates

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

STAT 13, UCLA, Ivo Dinov Slide 4

Questions

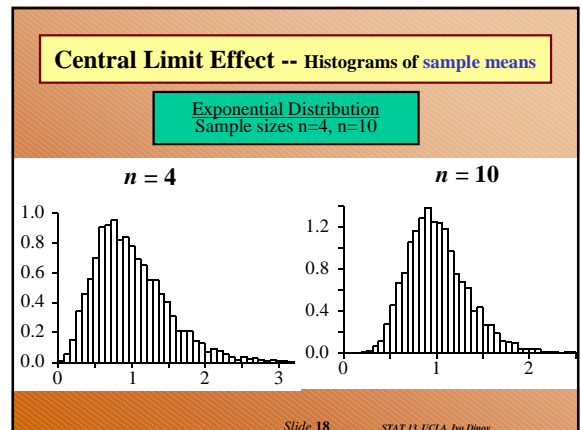
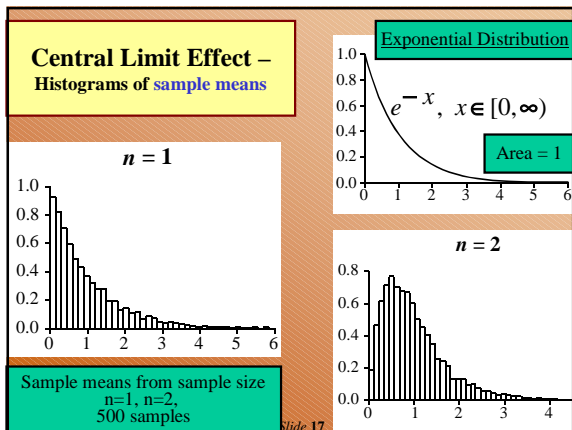
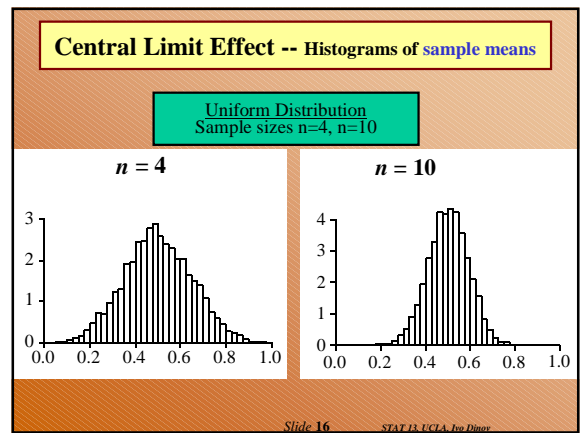
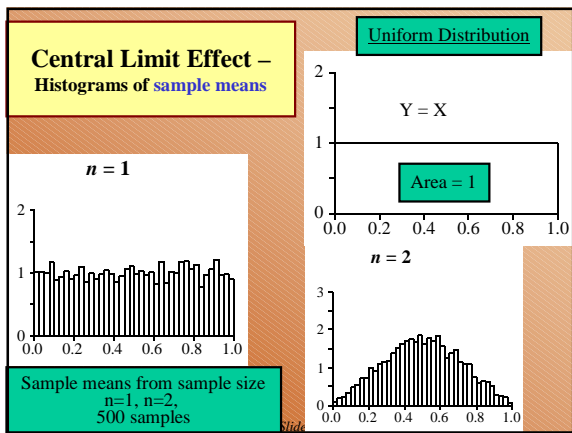
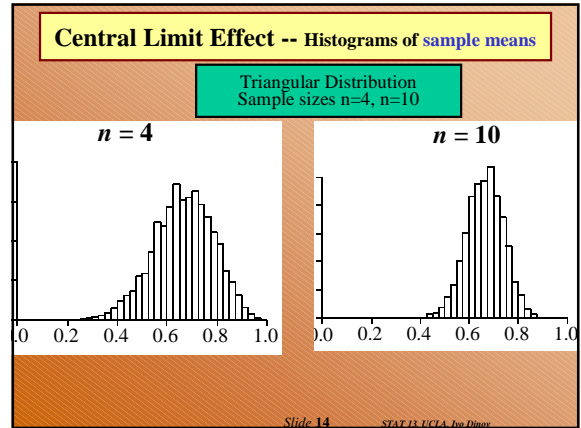
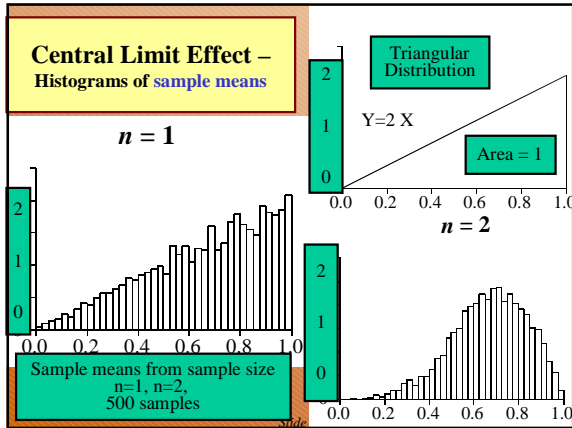
- What are two ways in which random observations arise and give examples. (random sampling from finite population – randomized scientific experiment; random process producing data.)
- What is a *parameter*? Give two examples of parameters. (characteristic of the data – mean, 1st quartile, std.dev.)
- 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 from 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^*)

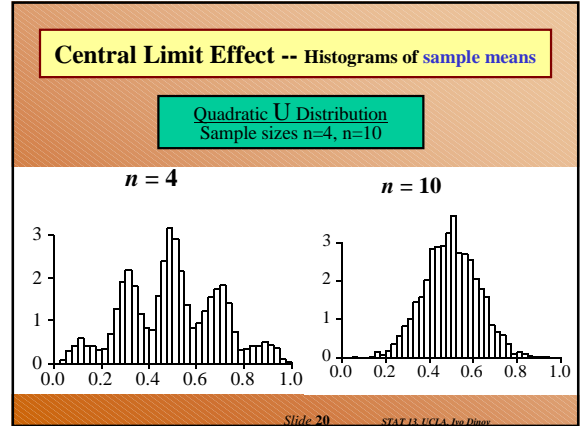
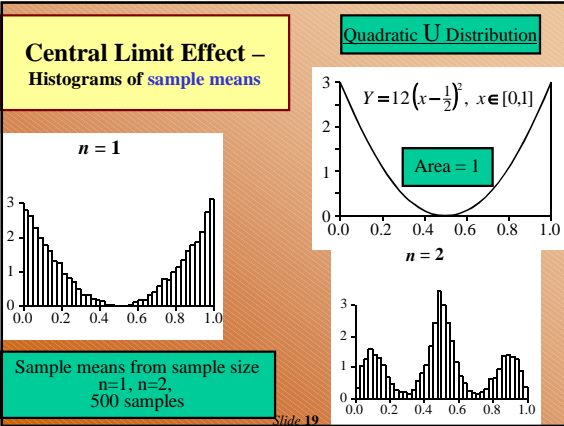
STAT 13, UCLA, Ivo Dinov Slide 5

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.

STAT 13, UCLA, Ivo Dinov Slide 6





Central Limit Theorem – heuristic formulation

Central Limit Theorem:
When sampling from almost any distribution, \bar{X} is approximately **Normally distributed** in large samples.

CLT Applet Demo
Applets.dir/SamplingDistributionApplet.html

Slide 21

Central Limit Theorem – theoretical formulation

Let $\{X_1, X_2, \dots, X_k, \dots\}$ 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 $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$, **sample-avg**,

Then \bar{X} has a **distribution** which approaches **$N(\mu, \sigma^2/n)$** , as $n \rightarrow \infty$.

Slide 22

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.)

Slide 23

Review

- When you have data from a moderate to small sample and want to use a **normal approximation** to the distribution of \bar{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** ($N(\mu, \sigma^2/n)$).

Slide 24

The standard error of the mean – remember ...

- 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*.
- Recall that for *known* $SD(X)=\sigma$, we can express the $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. How about if $SD(X)$ is *unknown*?!?

Slide 25 STAT 13, UCLA, Jon Dinger

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 $SE(\bar{x}) = \frac{\text{Sample standard deviation}}{\sqrt{\text{Sample size}}}$

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

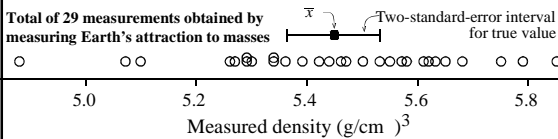
- Note similarity with
- $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Slide 26 STAT 13, UCLA, Jon Dinger

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H₂O

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 !

Slide 27 STAT 13, UCLA, Jon Dinger

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H₂O

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

and sample SD = $s_x = 0.2209457 \text{ g/cm}^3$

Then the standard error for these data is:

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

Slide 28 STAT 13, UCLA, Jon Dinger

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H₂O

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!

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

Slide 29 STAT 13, UCLA, Jon Dinger

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))

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

Slide 30 STAT 13, UCLA, Jon Dinger

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)

Slide 31 STAT 13, UCLA, Jon Dineen

Sampling distribution of the sample proportion

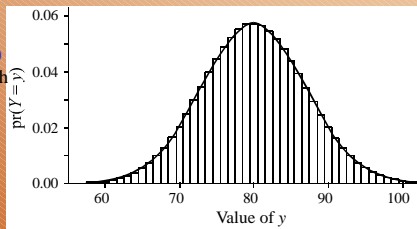
The sample proportion \hat{p} estimates the population proportion p .

Suppose, we poll college athletes to see what percentage are using performance inducing 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?

Slide 32 STAT 13, UCLA, Jon Dineen

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for $Y \sim \text{Bin}(n, p)$
 $\mu_Y = E(Y) = np$
 $\sigma_Y = SD(Y) = \sqrt{np(1-p)}$



For large samples, the distribution of \hat{P} is approximately Normal with

$$\text{mean} = p \quad \text{and} \quad \text{standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

Slide 33 STAT 13, UCLA, Jon Dineen

Approximate Normality in large samples

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

$$\begin{aligned} \mu_Y = E(Y) &= np & \mu_Z = E(Z) &= \frac{1}{n}E(Y) = p \\ \sigma_Y = SD(Y) &= \sqrt{np(1-p)} & \sigma_Z = SD(Z) &= \frac{1}{n}SD(Y) = \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

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

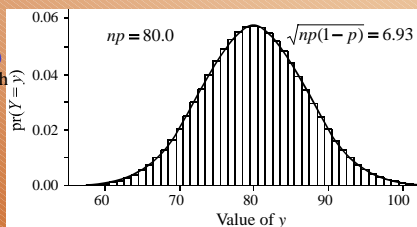
$$\mu_Z \pm 2SE(Z) = p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

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

Slide 34 STAT 13, UCLA, Jon Dineen

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for $Y \sim \text{Bin}(n, p)$



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 $N(np, \sqrt{np(1-p)})$

Slide 35 STAT 13, UCLA, Jon Dineen

Standard error of the sample proportion

Standard error of the sample proportion:

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

Slide 36 STAT 13, UCLA, Jon Dineen

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 $Y \sim \text{Bin}(n, p)$, when sample $< 10\%$ of popul.; $Y/n \sim \text{Normal}(m, s)$, 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?

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

Slide 37 STAT 13, UCLA, Jon Dineen

Review

- Why is the standard deviation of \hat{P} not useful in practice as a measure of the precision of the estimate?
 $SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$, 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

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

Slide 38 STAT 13, UCLA, Jon Dineen

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)?)

Slide 39 STAT 13, UCLA, Jon Dineen

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)

Slide 40 STAT 13, UCLA, Jon Dineen

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).

Slide 41 STAT 13, UCLA, Jon Dineen

Bias and Precision

- The **bias** in an estimator is the distance between the center of the sampling distribution of the estimator and the true value of the parameter being estimated. In math terms, $\text{bias} = E(\hat{\Theta}) - \theta$, where $\hat{\Theta}$ 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 $\frac{3}{4}$ of the sample mean?

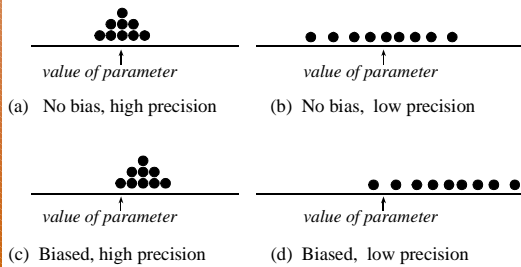
$$E(\hat{\Theta}) - \mu = E\left(\frac{1}{n} \sum_{k=1}^n X_k\right) - \mu = 0$$

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

Slide 42 STAT 13, UCLA, Jon Dineen

Bias and Precision

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



Slide 43 STAT 13, UCLA, Jon Dineen

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 $\hat{\theta}$.

Slide 44 STAT 13, UCLA, Jon Dineen

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**?

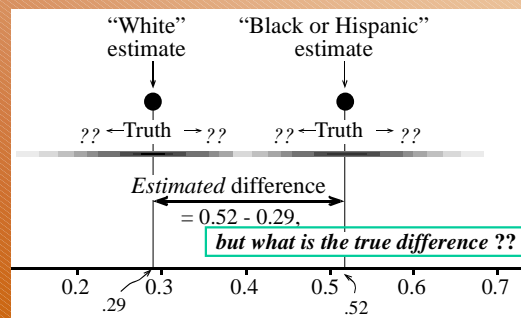
Slide 45 STAT 13, UCLA, Jon Dineen

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? ($10 \times 95 = 950$)

Slide 46 STAT 13, UCLA, Jon Dineen

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



Slide 47 STAT 13, UCLA, Jon Dineen

Standard error of a difference

Standard error for a difference between independent estimates:

$$SE(\text{Est}_1 - \text{Est}_2) = \sqrt{SE(\text{Est}_1)^2 + SE(\text{Est}_2)^2}$$

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

$N_1=139, p_1^{\wedge}=0.52, N_2=378, p_2^{\wedge}=0.29,$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.52 \times 0.48}{139} + \frac{0.29 \times 0.71}{378}} = 0.04838$$

Slide 48 STAT 13, UCLA, Jon Dineen

Standard error of a difference of proportions

Standard error for a difference between independent estimates:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.52 \times 0.48}{139} + \frac{0.29 \times 0.71}{378}} = 0.04838$$

So the estimated difference give/take 2SE's is:

$$\hat{p}_1 - \hat{p}_2 \pm 2 \times 0.04838 = [0.13; 0.33]$$

Slide 49 STAT 13, UCLA, Jon Dineen

Student's *t*-distribution

- For random samples from a Normal distribution,

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

Recall that for samples from $N(\mu, \sigma)$

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

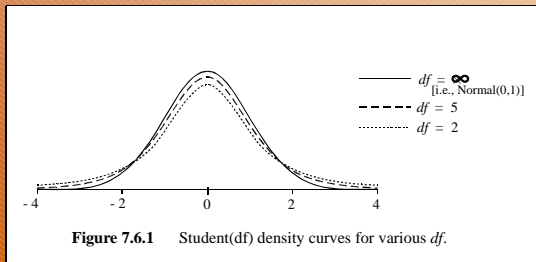
is **exactly** distributed as Student($df = n - 1$) ← Approx/Exact Distributions ↑

- 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.

Slide 50 STAT 13, UCLA, Jon Dineen

Density curves for Student's *t*

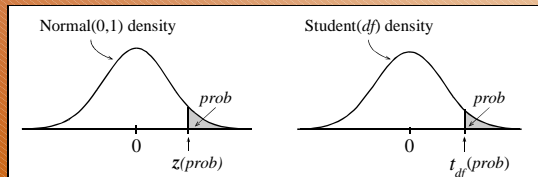


From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Slide 51 STAT 13, UCLA, Jon Dineen

Notation

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



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

Slide 52 STAT 13, UCLA, Jon Dineen

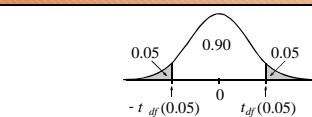


Figure 7.6.3 The central 90% of the Student(df) distribution.

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

Slide 53 STAT 13, UCLA, Jon Dineen

Reading Student's *t* table

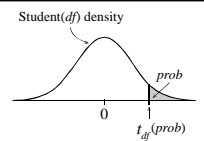


TABLE 7.6.1 Extracts from the Student's *t*-Distribution Table

df	prob									
	.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	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

Desired upper-tail prob

t -value

Slide 54 STAT 13, UCLA, Jon Dineen

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 $n \rightarrow \infty$)

Slide 55 STAT 13, UCLA, Jon Dinger

Review

- Why is T , the number of standard errors separating \bar{X} and μ , a more variable quantity than Z , the number of standard deviations separating \bar{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 $\bar{x} \pm 2 \text{se}(\bar{x})$ for a little more than 95% of all samples taken. For small samples from a normal distribution, is the proportion of samples for which the true value of μ lies within the 2-standard-error interval smaller or bigger than 95%? Why? (smaller – wider tail.)

Slide 56 STAT 13, UCLA, Jon Dinger

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

Slide 57 STAT 13, UCLA, Jon Dinger

Chapter 7 Summary

STAT 13, UCLA, Jon Dinger

Slide 58

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 , \bar{X} and \bar{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.

Slide 59 STAT 13, UCLA, Jon Dinger

Sampling distribution of \bar{X}

Sample mean, \bar{X} :

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

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

■ If we are sampling from a Normal distribution, then $\bar{X} \sim \text{Normal}$. (exactly)

■ **Central Limit Theorem:** For almost any distribution, \bar{X} is **approximately** Normally distributed in large samples.

Slide 60 STAT 13, UCLA, Jon Dinger

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:

- $\mu_p = E(\hat{P}) = p$ $\sigma_p = sd(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$

- \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)

Slide 61 STAT 13, UCLA, Joe Dimez

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 $E(\hat{\theta}) = \theta$.

Slide 62 STAT 13, UCLA, Joe Dimez

Precision

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

Slide 63 STAT 13, UCLA, Joe Dimez

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 \bar{x} is an unbiased estimate of the population mean μ
 - $SE(\bar{x}) = \frac{s_x}{\sqrt{n}}$

Slide 64 STAT 13, UCLA, Joe Dimez

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,

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

Slide 65 STAT 13, UCLA, Joe Dimez

TABLE 7.7.1 Some Parameters and Their Estimates

	Population(s) or Distributions(s) ↓ Parameters	Sample data ↓ Estimates	→ Measure of precision
Mean	μ	\bar{x}	$se(\bar{x})$
Proportion	p	\hat{p}	$se(\hat{p})$
Difference in means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$se(\bar{x}_1 - \bar{x}_2)$
Difference in proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$se(\hat{p}_1 - \hat{p}_2)$
General case	θ	$\hat{\theta}$	$se(\hat{\theta})$

Slide 66 STAT 13, UCLA, Joe Dimez

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.

Slide 67 STAT 13, UCLA, Joe Dimez

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 $T \sim \text{Student}(df)$, $\text{pr}(T \geq t) = prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.

Slide 68 STAT 13, UCLA, Joe Dimez

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

- If I ask 30 of you the question “Is 5 credit hour a reasonable load for Stat13?”, 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 13 format.
- $$\hat{p} \pm 2 \times SE(\hat{p}) = 0.5 \pm 2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.5 \pm 0.18$$
- If I ask all 300 Stat 13 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 ...

Slide 69 STAT 13, UCLA, Joe Dimez

T-distribution Examples

- Here are a couple of examples of using T-distribution that everyone should be familiar with:
 - Example 1: Suppose we have the following data $\{X_1=-1, X_2=2, X_3=0, X_4=1\}$ representing the results from a Test (T) on one population. Assuming only the test scores have normal distribution what is the chance the average, in general, exceeds 1.5?
 - Calculate the sample statistics
- $$\bar{X} = \sum_{i=1}^4 X_i = 0.5, \quad \sigma^2 = \frac{1}{4-1} \sum_{i=1}^4 (X_i - \bar{X})^2 = \frac{5}{6}, \quad \sigma = 0.9, \quad \frac{\sigma}{\sqrt{4}} = 0.45$$
- $$P(\bar{X} \geq 1.5) = (\text{standardize } 1.5) = P\left(T \geq \frac{1.5 - 0.5}{0.45}\right) = P(T \geq 2.22) = 0.054,$$
- where $T \sim t_{df=3}$, using the SOCR resource.

Slide 70 STAT 13, UCLA, Joe Dimez

T-distribution Examples

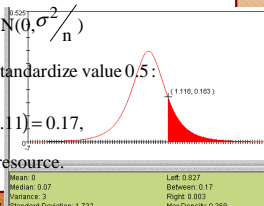
- Example 2: Suppose we have the same test data (T) as before but it represents a specific patient-population (e.g., cancer subjects) $\{X_1=-1, X_2=2, X_3=0, X_4=1\}$. Assuming the test scores have normal distribution and for normal individuals the mean response on the test T is zero. How atypical is the data? Could this test T be used as a patient marker?

Assume $X_i \sim N(\mu = 0, \sigma^2) \Rightarrow \bar{X} \sim N(0, \frac{\sigma^2}{n})$

$\bar{X} = \sum_{i=1}^4 X_i = 0.5, \quad \frac{\sigma}{\sqrt{4}} = 0.45$. Standardize value 0.5:

$$T = \frac{\bar{X} - \mu}{\sigma / \sqrt{4}} = \frac{0.5 - 0}{0.45} = 1.11, \quad P(T \geq 1.11) = 0.17,$$

where $T \sim t_{df=3}$, using the SOCR resource.



Slide