UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

•Instructor: Ivo Dinov,

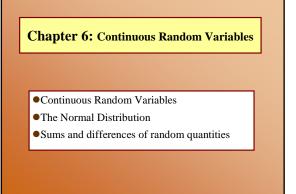
Asst. Prof. of Statistics and Neurology

• Teaching Assistants: Chris Barr & Ming Zheng

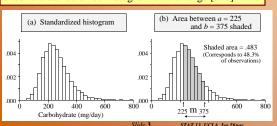
University of California, Los Angeles, Fall 2004 http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13. UCLA. Ivo Dinov

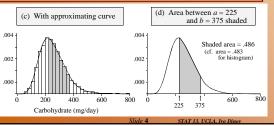
Slide

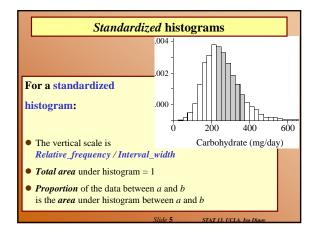


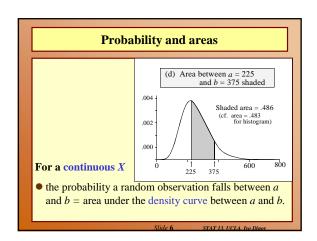
Dietary intake of carbohydrate (mg/day) for N=5929 people from a variety of work environments. Standardized histogram plot is unimodal but skewed to the right (high values). Vertical scale is (relative freq.)/(interval width) = $f_j/(N*m)$. The proportion of the data in [a : b] is the area under the standardized histogram on the range [a: b].

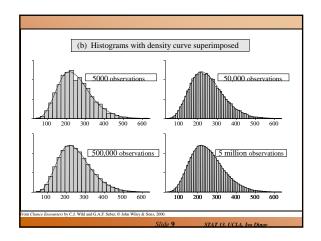


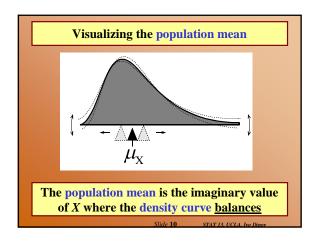
Superposition of a smooth curve (density function) on the standardized histogram (left panel). Area under the density curve on [a: b] = [225: 375] is analytically computed to be: 0.486 (right panel), which is close to the empirically obtained estimate of the area under the histogram on the same interval: 0.483 (left panel).











Interval endpoints and continuous variables

Recall a continuous variable is one where the <u>domain has</u> no gaps in between the values the variable can take.

In calculations involving a continuous random variable we *do not have to worry* about whether <u>interval</u> endpoints are included or excluded.

Slide 11 STAT 13. UCLA. Ivo Dinov

Review

- How does a standardized histogram differ from a relative-frequency histogram? raw histogram? (f/mm)
- What graphic feature conveys the proportion of the data falling into a class interval for a standardized histogram? for a relative-frequency histogram?
 (area=width . height = m f/mn=f/n)
- What are the <u>two</u> fundamental <u>ways</u> in which random observations arise? (Natural phenomena, sampling experiments – choose a student at random and use the lottery method to record characteristics, scientific experiments – blood pressure measure)
- How does a density curve describe probabilities?
 (The probability that a random obs. falls in [a:b] is the area under the PDF on the same interval.)

Slide 12 STAT 13. UCLA. Iva Dinas

Review

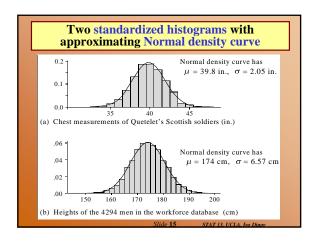
- What is the <u>total area</u> under both a standardized histogram and a probability density curve? (1)
- When can histograms of data from a random process be relied on to <u>closely resemble</u> the <u>density curve</u> for that process? (large sample size, small histogram bin-size)
- What characteristic of the density curve does the mean correspond to? (imaginary value of X, where the density curve balances)

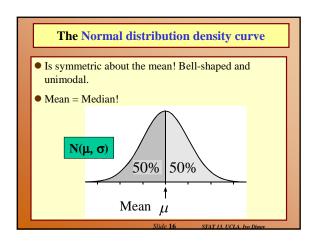
Slide 13 STAT 13 UCLA Iva Dina

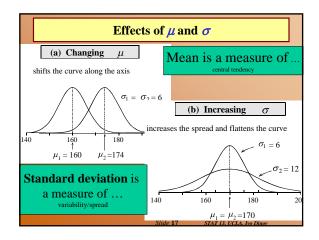
Review

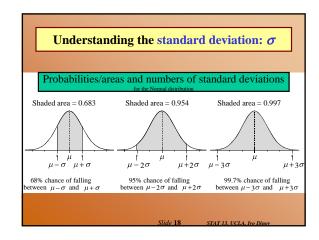
- Does it matter whether <u>interval endpoints</u> are included or excluded when we calculate probabilities for a continuous random variable from the area? (No)
- Why? (Area[a:b] == Area(a:b))
- Are discrete variables the <u>same</u> or <u>different</u> in this regard, interval endpoint not effecting the area? (Different)

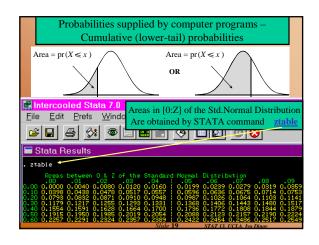
Slide 14 STAT 13. UCLA. Ivo Dinov

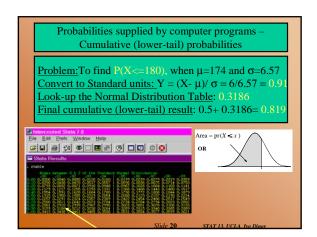






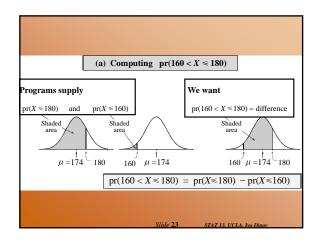


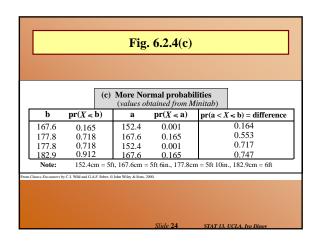


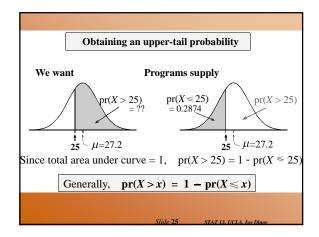


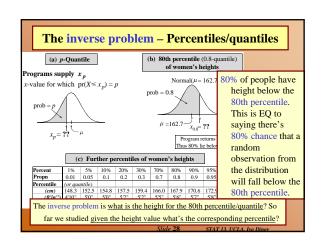


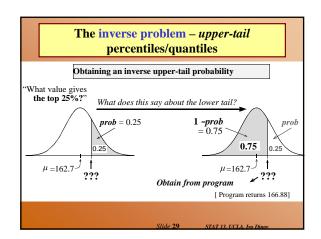
Sketch a Normal curve, marking the mean and other values of interest. Shade the area under the curve that gives the desired probability. Devise a way of getting the desired area from lowertail areas. Obtain component lower-tail probabilities from a computer program

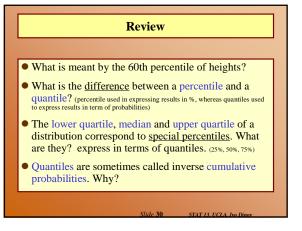


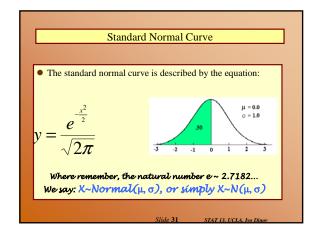


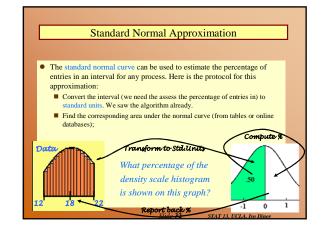


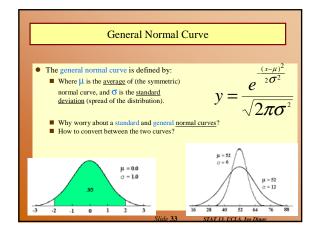


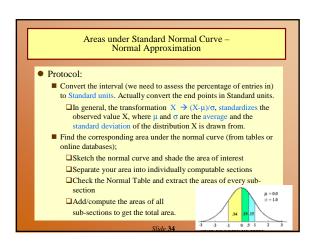


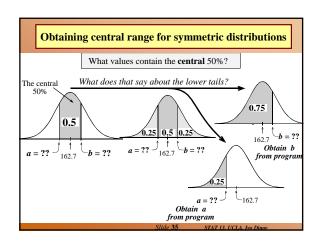


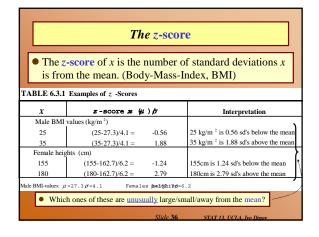


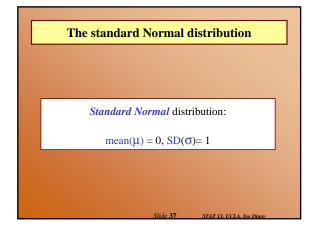


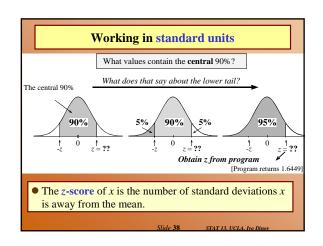






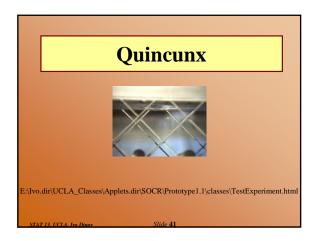






Working in standard units (# of SD's) TABLE 6.3.2 Central Ranges											
		Male BM	II values	Female heights							
Percentage	z	$\mu - z \sigma$	$\mu + z \sigma$	$\mu - z \sigma$	$\mu + z \sigma$						
80%	1.2816	22.05	32.55	154.8	170.6						
90%	1.6449	20.56	34.04	152.5	172.9						
95%	1.9600	19.26	35.34	150.5	174.9						
99%	2.5758	16.74	37.86	146.7	178.7						
99.9%	3.2905	13.81	40.79	142.3	183.1						
Male BMI-values	•	Inv	ndardizing erting	$Z = (X \circ X)$ $X = Z\sigma$	$-\mu)/\sigma$ $+\mu$						

TABLE	E6.3.3	Usin	gz-sc	ore tab	les							
As an ex	amp le,	we sha	all find	l pr(Z≤	≤ 1.135	7) usi	ng part	of the	table g	iven in		
Append	ix A4 (reprod	uced be	elow).								
Step 1:												
Step 2:	Look down the z column until you find 1.1. This tells you which row										w to	
	look ir	1.										
Step 3:	The second decimal place, here 4, tells you which column to look in.											
Step 4:	The er	The entry in the table corresponding to that row and column is										
	$pr(Z \le 1.14) = 0.873$											
						<u> </u>						
	z	0	1	2	3	4	5	6	7	8	9	
	1.0	.841	.844	.846	.848	.851	.853	.855	.858	.860	.862	
→	1.1	.864	.867	.869	.871	.873	.875	.877	.879	.881	.883	
	1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.901	
	1.3	.903	.905	.907	.908	.910	.911	.913	.915	.916	.918	
	1.4	.919	.921	.922	.924	.925	.926	.928	.929	.931	.932	
					Sli	le 40	ST	AT 13. U	CLA. Ivo I	Dinov		



Continuous Variables and Density Curves

- There are no gaps between the values a continuous random variable can take.
- Random observations arise in two main ways: (i) by sampling populations; and (ii) by observing processes.

Clida 42 CTATA NOTA IN DISC

The density curve

- The probability distribution of a continuous variable is represented by a density curve.
 - *Probabilities* are represented by *areas under the curve*,

 □ the probability that a random observation falls between *a* and *b* equal to the area under the density curve between *a* and *b*.
 - The total area under the curve equals 1.
 - The population (or distribution) mean $\mu_X = E(X)$, is where the density curve balances.
 - When we calculate probabilities for a continuous random variable, it does not matter whether interval endpoints are included or excluded.

Slide 43 STAT 13. UCLA. Ivo Dinov

For any random variable X

• E(aX + b) = a E(X) + b and SD(aX + b) = |a| SD(X)

Slide 44 STAT 13, UCLA, Ivo Dino

The Normal distribution

 $X \sim \text{Normal}(\mu_x = \mu, \sigma_x = \sigma)$

Features of the Normal density curve:

- The curve is a symmetric bell-shape centered at μ .
- The standard deviation σ governs the spread.
 - 68.3% of the probability lies within 1 standard deviation of the mean
 - 95.4% within 2 standard deviations
 - 99.7% within 3 standard deviations

Slide 45 STAT 13 UCLA Ivo Dinor

Probabilities

- Computer programs provide lower-tail (or cumulative) probabilities of the form $pr(X \le x)$
 - We give the program the *x*-value; it gives us the probability.
- Computer programs also provide inverse lower-tail probabilities (or quantiles)
 - We give the program the probability; it gives us the *x*-value.
- When calculating probabilities, we shade the desired area under the curve and then devise a way of obtaining it via lower-tail probabilities.

Slide 46 STAT 13, UCLA, Ivo Dino:

Standard Units

The z-score of a value a is

- the number of standard deviations *a* is away from the mean
- positive if *a* is above the mean and negative if *a* is below the mean.

The *standard Normal* distribution has $\mu = 0$ and $\sigma = 0$.

 We usually use Z to represent a random variable with a standard Normal distribution.

Slide 47 STAT 13, UCLA, Ivo Dine

Ranges, extremes and z-scores

Central ranges:

■ $P(-z \le Z \le z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls within z SD's either side of the mean.

Extremes:

- $P(Z \ge z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls more than z standard deviations above the mean.
- $P(Z \le -z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls more than z standard deviations below the mean.

Slide 48 STAT 13. UCLA. Ivo Dinov

Combining Random Quantities

Variation and independence:

- No two animals, organisms, natural or man-made objects are ever identical.
- There is always variation. The only question is whether it is large enough to have a practical impact on what you are trying to achieve.
- Variation in component parts leads to even greater variation in the whole.

Slide 49 STAT 13. UCLA. Ivo Dinov

Independence

We model variables as being independent

- if we think they relate to physically independent processes
- and if we have no data that suggests they are related.

Both sums and differences of independent random variables are more variable than any of the component random variables

Slide 50 STAT 13. UCLA. Ivo Dinov

Formulas

- For a constant number a, E(aX) = aE(X) and SD(aX) = |a| SD(X).
- Means of sums and differences of random variables act in an obvious way
 - the mean of the sum is the sum of the means
 - the mean of the difference is the difference in the means
- For independent random variables, (cf. Pythagorean theorem), $SD(X_1 + X_2) = SD(X_1 X_2) = \sqrt{SD(X_1)^2 + SD(X_2)^2}$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

[ASIDE: Sums and differences of independent Normally distributed random variables are also Normally distributed]

Slide 51 STAT 13, UCLA, Ivo Dinov

Example

- Assumption: Crime rate for individuals is independent of family relations!
- Let X = RV representing the number of crimes an average individual commits in a 5 yr span.
- Let X1 and X2 be the crime rates of a husband and the wife in one family. What is the expected crime rate for this family given that E(X) = 1.4 and SD(X) = 0.7?

$$SD(X_1 + X_2) = SD(X_1 - X_2) = \sqrt{SD(X_1)^2 + SD(X_2)^2}$$

 $E(X_1 + X_2) = E(X_1) + E(X_2)$

Slide 52 STAT 13, UCLA, Iyo Dino

Example

- Assumption: Crime rate for individuals is independent of family relations!
- If X1+X2 = 3X-1, dependent case
- X1 and X2 are independent (family relation)
- Suppose X1+X2 = 5, is this atypical?

$$SD(X_1 + X_2) = SD(X_1 - X_2) = \sqrt{SD(X_1)^2 + SD(X_2)^2}$$

 $E(X_1 + X_2) = E(X_1) + E(X_2)$

Slide 53 STAT 13, UCLA, Ivo Dine

Percentage is 38.30

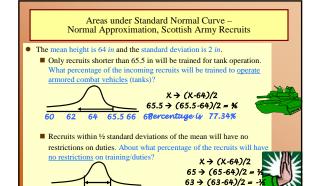
Areas under Standard Normal Curve – Normal Approximation

- Protocol:
 - Convert the interval (we need to assess the percentage of entries in) to Standard units. Actually convert the end points in Standard units.
 - □ In general, the transformation $X \rightarrow (X-\mu)/\sigma$, standardizes the observed value X, where μ and σ are the average and the standard deviation of the distribution X is drawn from.
 - Find the corresponding area under the normal curve (from tables or online databases);
 - ■Sketch the normal curve and shade the area of interest
 - ☐ Separate your area into individually computable sections
 - □ Check the Normal Table and extract the areas of every sub-
 - ☐Add/compute the areas of all

section

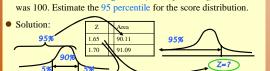
sub-sections to get the total area

a. 34 19 15



62 63 64 65 66 68

When the histogram of the observed process follows the normal curve Normal Tables (of any type, as described before) may be used to estimate percentiles. The N-th percentile of a distribution is P is N% of the population observations are less than or equal to P. Example, suppose the Math-part SAT scores of newly admitted freshmen at UCLA averaged 535 (out of [200:800]) and the SD



Percentiles for Standard Normal Curve • Example, suppose the Math-part SAT scores of newly admitted freshmen at UCLA averaged 535 (out of [200:800]) and the SD was 100. Estimate the 95 percentile for the score distribution. • Solution: 2 Area 1.65 90.11 1.70 91.09 95% • Z=1.65 (std. Units) \Rightarrow 700 (data units), since $X \Rightarrow (X - \mu)/\sigma$, converts data to standard units and $X \Rightarrow \sigma X + \mu$, converts standard to data units! $\sigma = 100$: $\mu = 535$, $100 \times 1.65 + 535$ 700.

Summary

- 1. The Standard Normal curve is symmetric w.r.t. the origin (0,0) and the total area under the curve is 100% (1 unit)
- 2. Std units indicate how many SD's is a value below (-)/above (+) the
- Many histograms have roughly the shape of the normal curve (bell-shape)
- If a list of numbers follows the normal curve the percentage of entries falling within each interval is estimated by: 1. Converting the interval to StdUnits and, 2. Computing the corresponding area under the normal curve (Normal approximation)
- 5. A histogram which follows the normal curve may be reconstructed just from $(\mu,\sigma^2),$ mean and variance=std_dev^2
- 6. Any histogram can be summarized using percentiles
- E(aX+b)=aE(X)+b, Var(aX+b)=a²Var(X), where E(Y) the the mean of Y and Var(Y) is the square of the StdDev(Y),

Slide 58 STAT 13, UCLA, Ivo Dinov

