STA 570 Spring 2011

Lecture 8 *Thursday, Feb 10*

- **Summarizing Bivariate Data**
  - **Two quantitative variables: Least squares regression**
- **Normal Distribution**
- **z-Scores**

Homework 5: Due next week in lab.

Review: Method of Least Squares (Gauss)

- Minimize the sum of the squared residuals

$$\sum (y_i - \hat{y}_i)^2$$

- The squared residuals are the squared vertical distances between the straight line and the data

- ***Prediction equation*** or ***least squares equation***

$$\hat{y} = b_0 + b_1 \cdot x$$

Review: Correlation and Regression

- ***The correlation coefficient r*** measures the ***strength and direction of the linear association*** between *X* and *Y*
- r is always between –1 and +1
- It is not affected by (linear) unit changes or by switching the roles of explanatory (*x*) and dependent (*y*) variable

- ***The slope of the prediction equation*** provides the expected change in *y* (rise) for a one-unit increase in *x* (run)
- It is affected by unit changes, and it changes when the roles of *x* and *y* are switched
- The intercept of the prediction equation is the (hypothetical) predicted value of *y* for *x=0*
- It often has little practical meaning

Some Common Mistakes or Misuses of Correlation and Regression

- Only reporting the numerical values of correlation coefficient (*r*) or prediction equation, without a scatter plot.
- Using correlation or linear regression for data that shows a clearly nonlinear association between the two variables.
- Claiming *no association* when in fact there is *no linear association.*
- Inferring *causation* from *association.*
- The bivariate sample is not a random sample of the population. In particular, the *x*-values in the sample are not representative of the *x*-values in the population.
- Extrapolation

Always use common sense…!

Effect of Outliers

- Outliers can have a substantial effect on the (estimated) prediction equation
- In the murder rate vs. poverty rate example, DC is an outlier
- Prediction equation with DC:

  y-hat = -10.13 + 1.32 x
- Prediction equation without DC:

  y-hat = -0.86 + 0.58 x

Effect of Outliers

- Removing the outlier would cause a large change in the results
- Observations whose removal causes substantial changes in the prediction equation, are called ***influential***
- It may be better not to use one single prediction equation for the 50 states and DC
- In reporting the results, it has to be noted that the outlier DC has been removed
- Correlation and Regression Applet

Model Assumptions and Violations

- **Influential Observations**
- Main disadvantage of least squares method: It is not robust against the effect of influential observations
- One single observation can have a large effect on the prediction equation
  - if its *X* value is unusually large or small,
  - and if its *Y* value falls far from the trend that the rest of the data follow

Prediction

- The prediction equation $y\text{-}hat = b_0 + b_1 x$ is used for predictions about the response variable *y* for different values of the explanatory variable *x*
- For example, based on the poverty rate, the predicted murder rate for Arizona is

$$b_0 + b_1 x = -0.8567 + 0.5842 \times 20 = 10.83$$

Dependent Variable	Predicted Value	Residual	
10.2	10.8281	-0.6281	*(Arizona)*
6.6	11.0618	-4.4618	*(Kentucky)*

Residuals

- The difference between observed and predicted values of the response variable (*y* – *y-hat*) is called a ***residual***
- The residual is negative when the observed value is smaller than the predicted value
- The smaller the absolute value of the residual, the better is the prediction
- The sum of all residuals is zero

Scatterplot

- Is linear regression/correlation always appropriate for two quantitative variables?
- How to decide whether a linear function may be used?
- *Always **plot the data** first*
- Recall: A ***scatterplot*** is a plot of the values (x,y) of the two variables
- Each subject is represented by a point in the plot
- If the plot reveals a non-linear relation, then linear regression is not appropriate, and the (Pearson) correlation coefficient is not informative

Model Assumptions and Violations

- **Models and Reality**
- The regression model only *approximates* the true relationship between two variables.
- In practice, these relationships are rarely exactly linear.
- Sometimes, the simple linear regression is too simplistic, and a more general model needs to be found.
- **A good regression model is realistic and describes the relationship adequately, but is still simple enough to be easily interpreted.**

The Normal (*Gaussian, Bell Curve*) Distribution

- Carl Friedrich Gauß (1777-1855), ***Gaussian Distribution***

- Normal distribution is perfectly ***symmetric*** and ***bell-shaped***
- Characterized by two parameters:

  mean μ and ***standard deviation σ***
- The ***68%-95%-99.7%*** rule applies "precisely"* to the normal distribution

*More precisely: 68.26895% - 95.44997% - 99.73002%


DEUTSCHE BUNDESBANK
Banknote
10
1777 1855 Carl Friedr. Gauß
ZEHN DEUTSCHE MARK
DN6833178D9

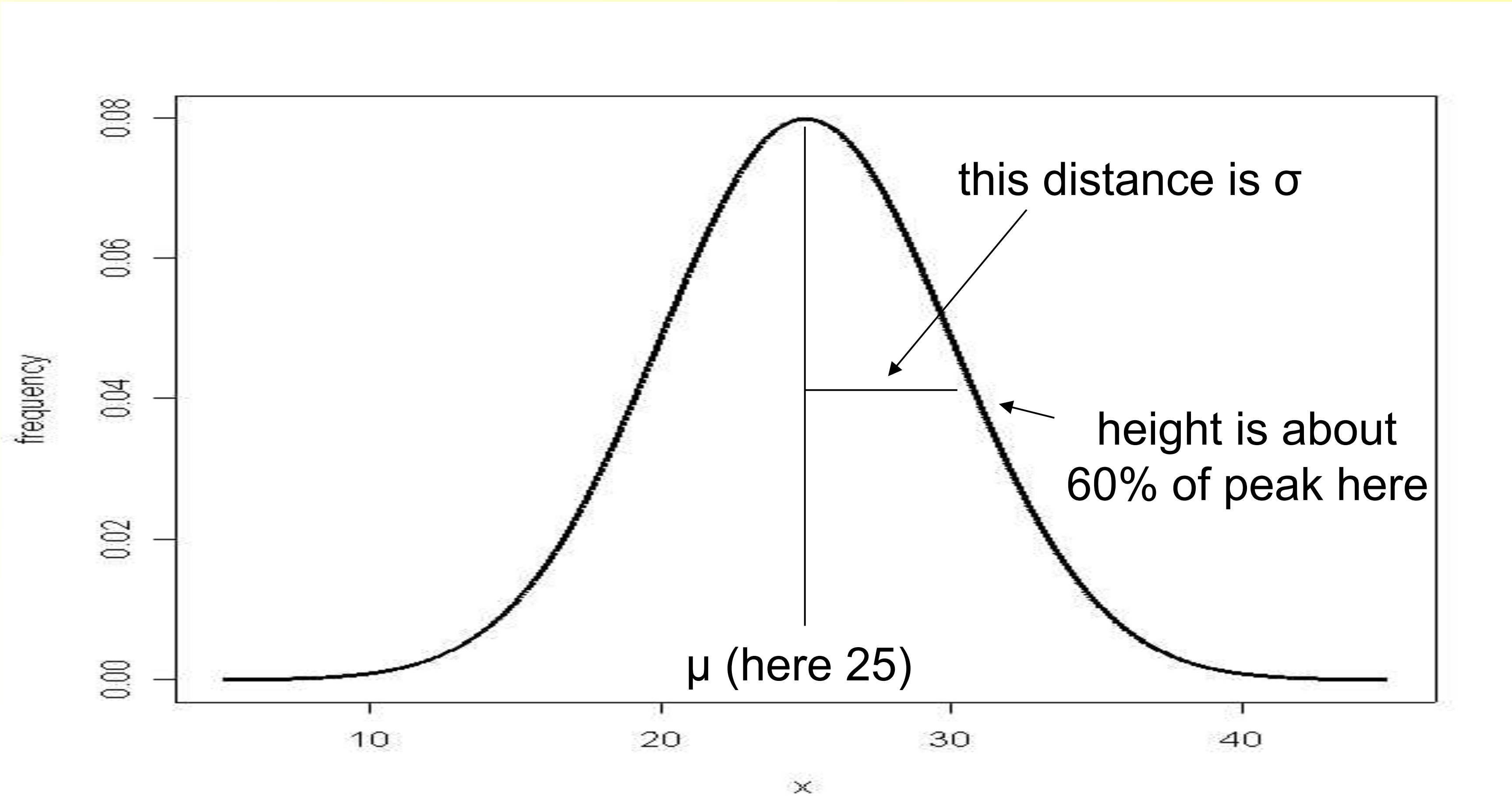
The Normal Distribution is Common

- Many real data follow a normal shape. For example
- 1) Many/most biometric measurements (heights, femur lengths, skull diameters, etc.)
- 2) Scores on many standardized exams (IQ tests, SAT, ACT) are forced into a normal shape before reporting
- 3) Microarray expression intensities (if you take the logarithm first)
- 4) Averages of measurements!

Mean and Standard Deviation

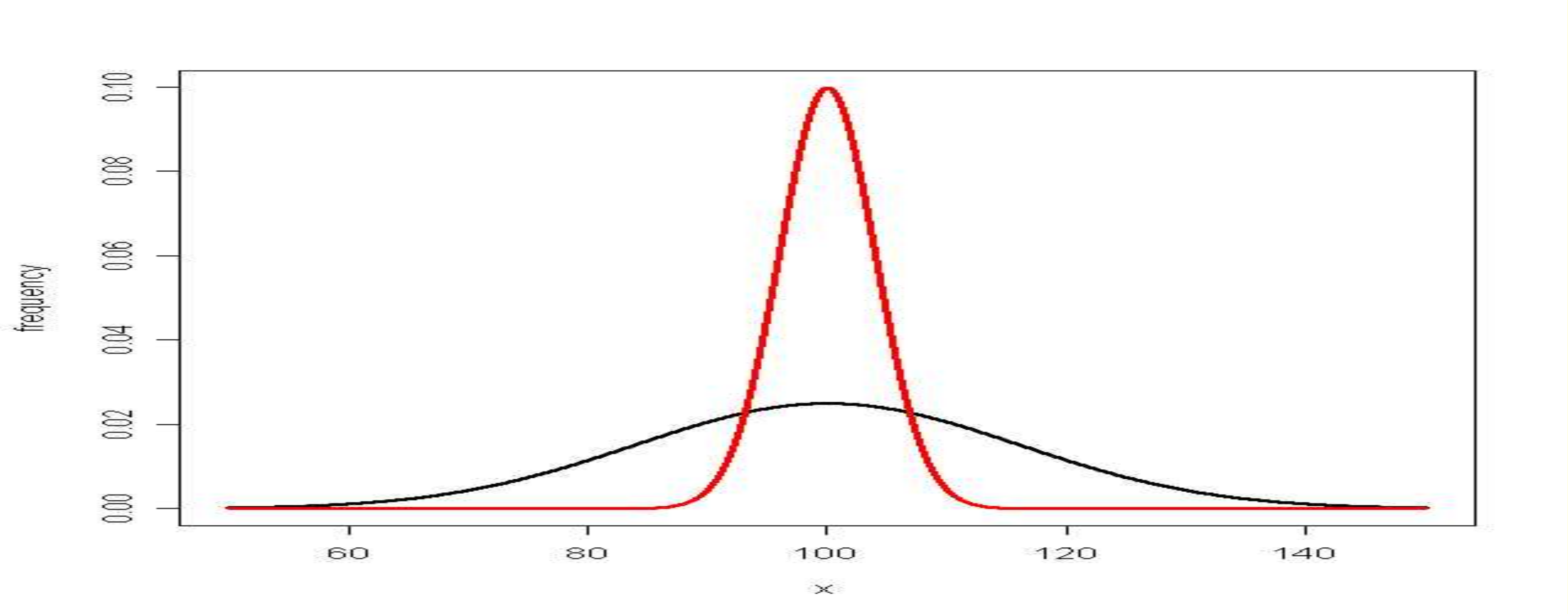
- Normal distributions are characterized by two numbers
- mean or “expected value” (corresponding to the peak)
- “standard deviation” (distance from mean to inflection point)
- Large standard deviations result in “spread out” normal distributions.
- Small standard deviations result in “strongly peaked” distributions.

Mean (μ) and Standard deviation (σ) for a normal distribution

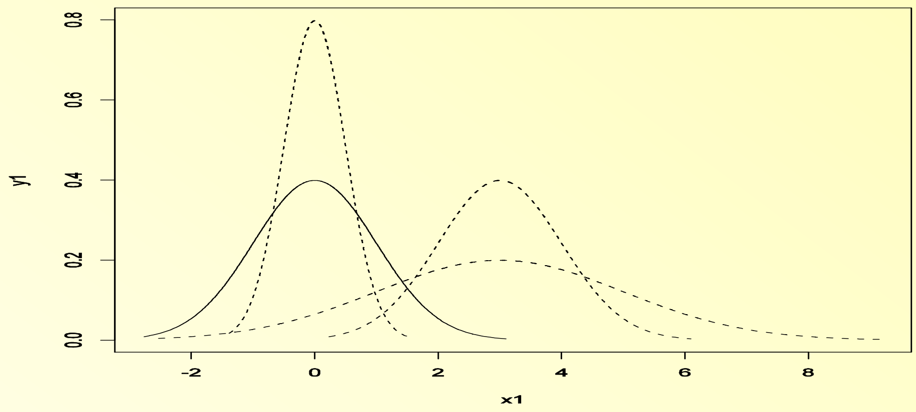


Two Normal Distributions, Corresponding to Different Standard Deviations

- Mean=100, std.dev = 16
- Mean=100, std.dev = 4



More Normal Distributions



Describing Normal Distributions

- Central location: mean μ (=median).
- Spread: standard deviation σ (interquartile range is about 4/3 σ)
- Shape: Normal distributions are symmetric and typically have few, if any, outliers.

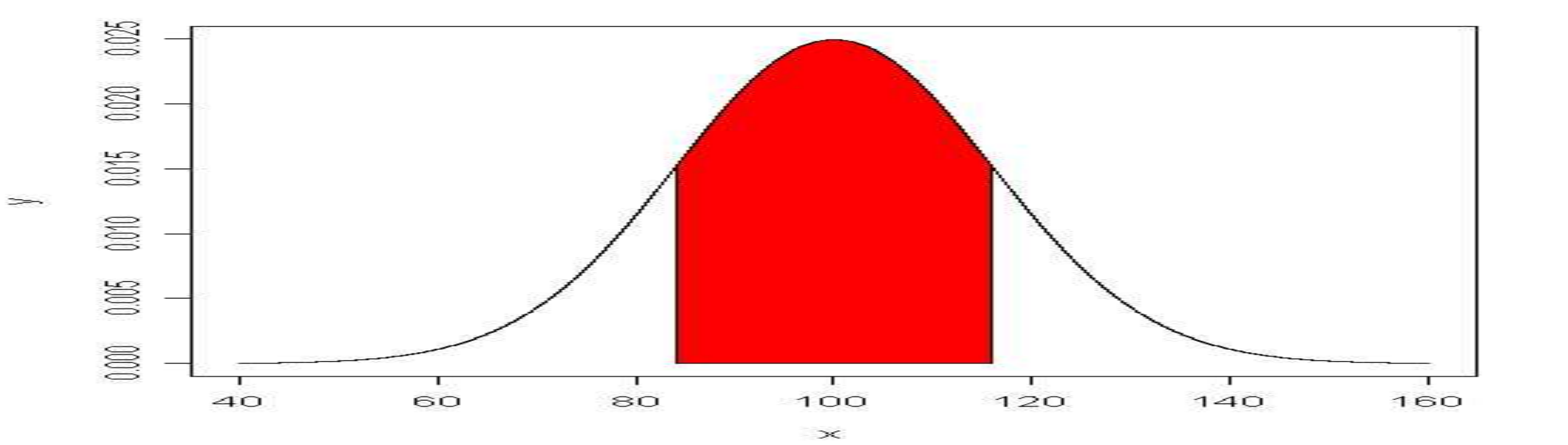
- If your data has a lot of outliers, but is otherwise symmetric and unimodal, it may have a “t” distribution (discussed later in class).

Probabilities from a Normal distribution

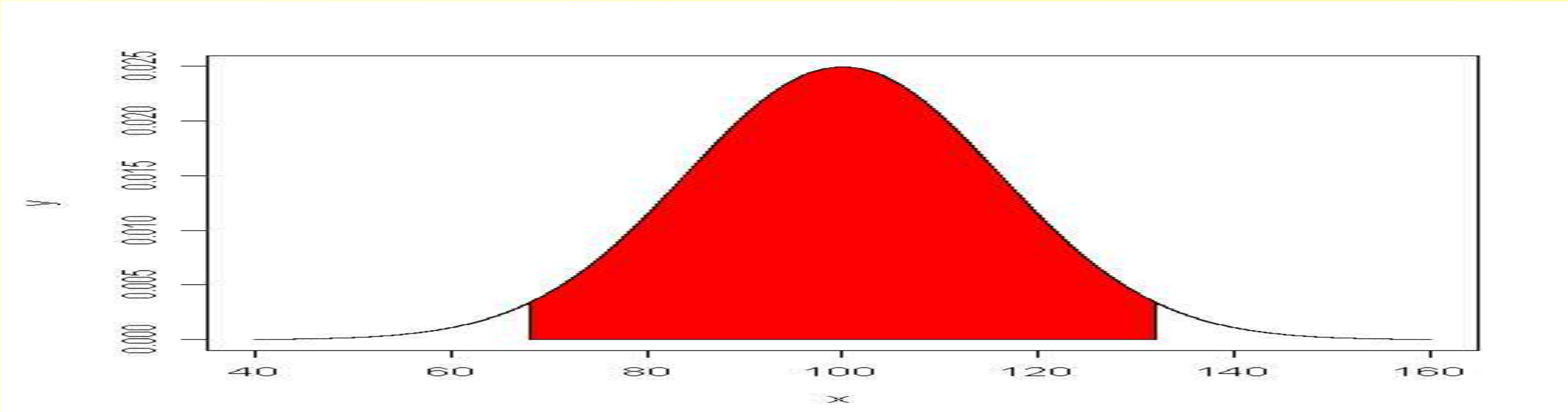
- Normal distributions have a nice property that, knowing the mean (μ) and standard deviation (σ), we can tell how much data will fall in any region.
- In other words: The complete distribution is determined by the two parameters.
- Examples – the normal distribution is symmetric, so 50% of the data is smaller than μ and 50% is larger than μ.

Verifying the Empirical Rule

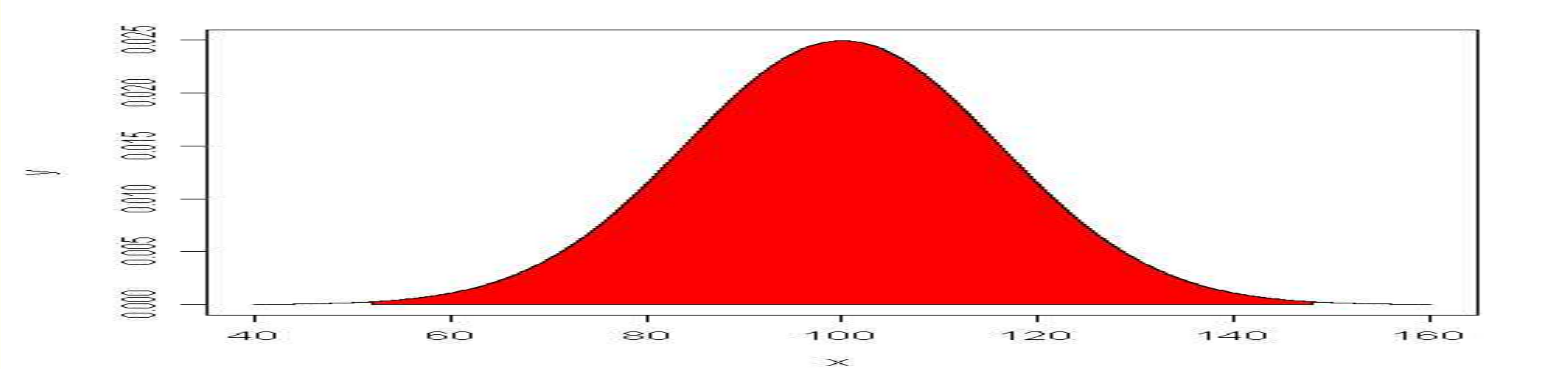
- It is always true that about 68% of the data appears within 1 standard deviation of the mean (so about 68% of the data appears in the region μ±σ)
- Normal Density Curve Applet



95% within 2 standard deviations



99.7% of the data (almost all the data) within 3 standard deviations of the mean



- In quality control applications, one often is interested in "6-sigma".
- 6 standard deviations include 99.9999998% of the data.

Normal Distribution: Example (female height)

- Assume that adult female height has a normal distribution with mean μ=165 cm and standard deviation σ=9 cm
- With probability 0.68, a randomly selected adult female has height between

  $\mu - \sigma = 156$ cm and $\mu + \sigma = 174$ cm
- With probability 0.95, a randomly selected adult female has height between

  $\mu - 2\sigma = 147$ cm and $\mu + 2\sigma = 183$ cm
- Only with probability 1-0.997=0.003, a randomly selected adult female has height below

  $\mu - 3\sigma = 138$ cm or above $\mu + 3\sigma = 192$ cm

Normal Distribution

- So far, we have looked at the probabilities within one, two, or three standard deviations from the mean

  ($\mu + \sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$)
- How much probability is concentrated within 1.43 standard deviations of the mean?
- More general, how much probability is concentrated within *z* standard deviations of the mean?

Normal Distribution Calculators

- Many statistics textbooks contain tables of the normal distribution probabilities
- Online tools are easier to use, and more precise
- Standard Normal Calculator "Surfstat"
- Standard Normal Calculator "Stat Trek”

- Example, for z=1.43:

The probability within 1.43 standard deviations of the mean is _____.

The probability outside 1.43 standard deviations of the mean is _____.

Working backwards

- We can also use the online calculator to find *z*-values for given probabilities
- Find the *z*-value corresponding to a right-hand tail probability of 0.025
- Answer: Probability 0.025 lies above

  μ + _____ σ
- Find the *z*-value for a right-hand tail probability of 0.1, 0.05, 0.01

Finding *z*-Values for Percentiles

- For a normal distribution, how many standard deviations from the mean is the 90th percentile?
- Or: What is the value of z such that 0.90 probability is less than $\mu + z\,\sigma$?
- Answer: The 90th percentile of a normal distribution is ______ standard deviations above the mean

Application

- SAT scores are approximately normally distributed with mean 500 and standard deviation 100
- The 90^{th} percentile of the SAT scores is 1.28 standard deviations above the mean
- $\mu + 1.28\,\sigma = 500 + 1.28 \cdot 100 = 628$
- Find the 99^{th} and the 5^{th} percentile of SAT scores

Online Tools

http://bcs.whfreeman.com/scc/content/cat_040/spt/normalcurve/normalcurve.html

http://stat.utilities.googlepages.com/tables.htm

http://stattrek.com/Tables/Normal.aspx

- Use these to
  - verify graphically the empirical rule,
  - find probabilities,
  - find percentiles
  - calculate *z*-values for one- and two-tailed probabilities

Example

- In baseball, batting average is the number of hits divided by the number of at-bats.
- Recent batting averages of almost 1000 Major League Baseball players could be described by a normal distribution with mean 0.270 and standard deviation 0.008.
- What percent of the players have a batting average of 0.28 and greater?
- What percentage have a batting average of below 0.25?
- If there are 30 players on a roster, how many would you expect to have a batting average of above 0.28 (below 0.25)

Another Example

- Assume that cholesterol levels of men in the US have an approximately normal distribution with mean 215 (mg/dl) and standard deviation 25 (mg/dl).
- What is the probability that the cholesterol level of a randomly selected man is less than 180?
- What is the probability that it is between 190 and 220?

Quartiles of Normal Distributions

- Median: z=0

  (0 standard deviations above the mean)
- Upper Quartile: z = 0.67

  (0.67 standard deviations above the mean)
- Lower Quartile: z= – 0.67

  (0.67 standard deviations below the mean)
- Find the lower and upper quartile of cholesterol levels for men in the US

z-Scores

- The *z*-score for a value x of a random variable is the number of standard deviations that x is above μ
- If x is below μ, then the *z*-score is negative
- The *z*-score is used to compare values from different normal distributions

Calculating *z*-Scores

- You need to know x, μ, and σ to calculate *z*

$$z = \frac{x - \mu}{\sigma}$$

Tail Probabilities

- SAT Scores: Mean=500,
  Standard Deviation =100
- The SAT score 700 has a *z*-score of *z*=2
- The probability that a score is ***beyond*** 700 is the tail probability of *z*=2
- Online tool….
- 2.28% of the SAT scores are ***beyond*** 700 (***above*** 700)

Tail Probabilities

- SAT score 450 has a *z*-score of *z*=-0.5
- The probability that a score is ***beyond*** 450 is the tail probability of *z*=-0.5
- Online tool….
- 30.85% of the SAT scores are ***beyond*** 450 (***below*** 450)

z-Scores

- The z-score is used to compare values from different normal distributions
- SAT: μ=500, σ=100
- ACT: μ=21, σ=6
- What is better, 650 in the SAT or 28 in the ACT?

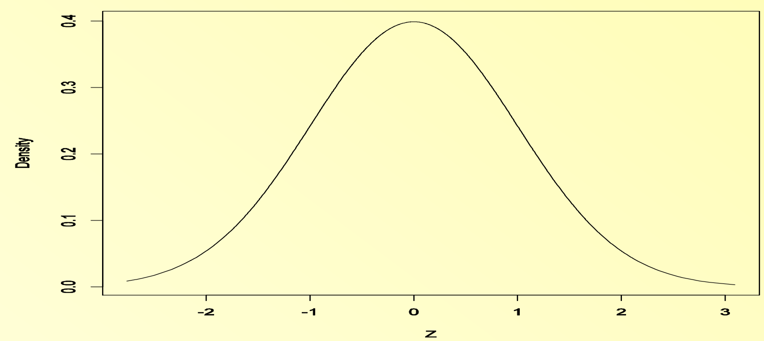
$$z_{SAT} = \frac{x-\mu}{\sigma} = \frac{650-500}{100} = 1.5$$

$$z_{ACT} = \frac{x-\mu}{\sigma} = \frac{28-21}{6} = 1.17$$

Corresponding tail probabilities?
How many percent have better
SAT or ACT scores?

Standard Normal Distribution

- The standard normal distribution is the normal distribution with mean μ=0 and standard deviation σ=1



Standard Normal Distribution

- When values from an arbitrary normal distribution are converted to *z*-scores, then they have a standard normal distribution
- The conversion is done by subtracting the mean μ, and then dividing by the standard deviation σ

$$z = \frac{x - \mu}{\sigma}$$