Inferential Statistics

PROF. DR. AKIN PALA

Inferential Statistics

- Statistical inference is the science of characterizing or making decisions about a population using information from a sample drawn from that population.
- Most of the practice of statistics is concerned with inferential statistics.
- Statistical inference is a refinement of normal inference, and is a process of making generalizations about unmeasured populations using data calculated on measured samples.

Inferential Statistics

- It is easy to confuse descriptive statistics and inferential statistics, because the methods are the same, but interpretation is different.
- μ for population mean is different than $\overline{X}$ for the sample mean, or N vs. n
  - but they are calculated the same way.
- Some have different formulas:
  - Population standard deviation we divide by *N*
  - *Sample standard deviation we* divide by $n - 1$.
- Any time you want to generalize your results beyond the specific cases that provided your data, you should be doing inferential statistics.

Pdf: Probability Density Function

- A theoretical probability distribution is defined by a formula that specifies
  - Discrete distributions: what values can be taken by data points within the distribution, and how common each value will be or,
  - Continuous distributions: how common a given range of values will be.
- Graphic presentations of theoretical probability distributions are used to present statistical concepts:
  - Well-known "bell curve" of the normal distribution is a good example

Pdf: Probability Density Function

- Probability distributions are
- *Continuous: the data* can take on any value within a specified range
  - The normal distribution is an example of a continuous distribution
- *Discrete: the data can* only take on certain values.
  - The binomial distribution is an example of a discrete distribution.

Pdf: Probability Density Function, Discrete

- Any variable has a density function
- The probability distribution of a discrete random variable lists all the possible values that the random variable can assume and their corresponding probabilities.
- The probability distribution of x describes how the probabilities are distributed over all the possible values of x.

Pdf: Probability Density Function, Discrete

Frequency and Relative Frequency Distributions of the Number of Vehicles Owned by Families		
Number of Vehicles Owned	**Frequency**	**Relative Frequency**
0	30	.015
1	470	.235
2	850	.425
3	490	.245
4	160	.080
	n=2000	Sum=1

Pdf: Probability Density Function, Discrete

Frequency and Relative Frequency Distributions of the Number of Vehicles Owned by Families	
Number of Vehicles Owned x	**Probability P(x)**
0	.015
1	.235
2	.425
3	.245
4	.080
	Sum of all probabilities=1

Pdf: Probability Density Function, continuous

- A probability density function (PDF), or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value.
- Suppose a species of bacteria typically lives 4 to 6 hours. What is the probability that a bacterium lives *exactly* 5 hours?
  - The answer is 0%. Lots of bacteria live for *approximately* 5 hours, but there is negligible chance that any given bacterium dies at *exactly* 5.0000000000... hours.

Pdf: Probability Density Function, continuous

- What is the probability that the bacterium dies between 5 hours and 5.01 hours?
  - Let's say the answer is 0.02 (i.e., 2%).
- What is the probability that the bacterium dies between 5 hours and 5.001 hours?
  - The answer is probably around 0.002, since this is 1/10th of the previous interval.
  - The probability that the bacterium dies between 5 hours and 5.0001 hours is probably about 0.0002 etc.

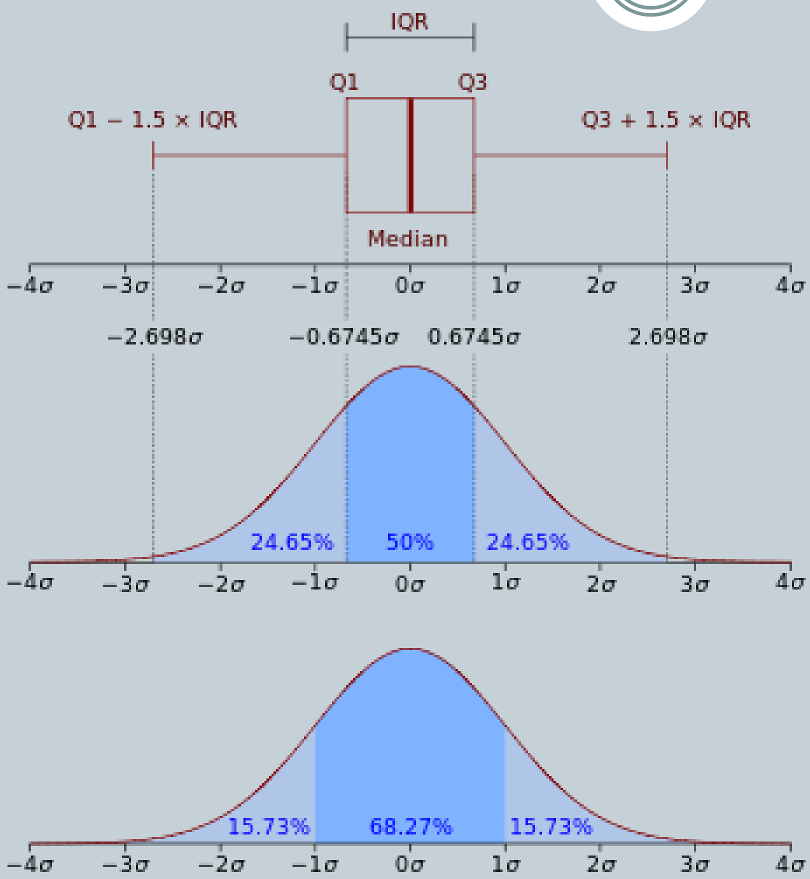
Pdf: Probability Density Function, continuous

- The ratio :

$$\frac{\text{probability of dying during an interval}}{\text{duration of the interval}}$$

- is approximately constant, and equal to 2 per hour (or 2 hour^{-1}).
- There is 0.02 probability of dying in the 0.01-hour interval between 5 and 5.01 hours, and (0.02 probability / 0.01 hours) = 2 hour^{-1}.
- This quantity 2 hour^{-1} is called the *probability density* for dying at around 5 hours.

Pdf: Probability Density Function, continuous



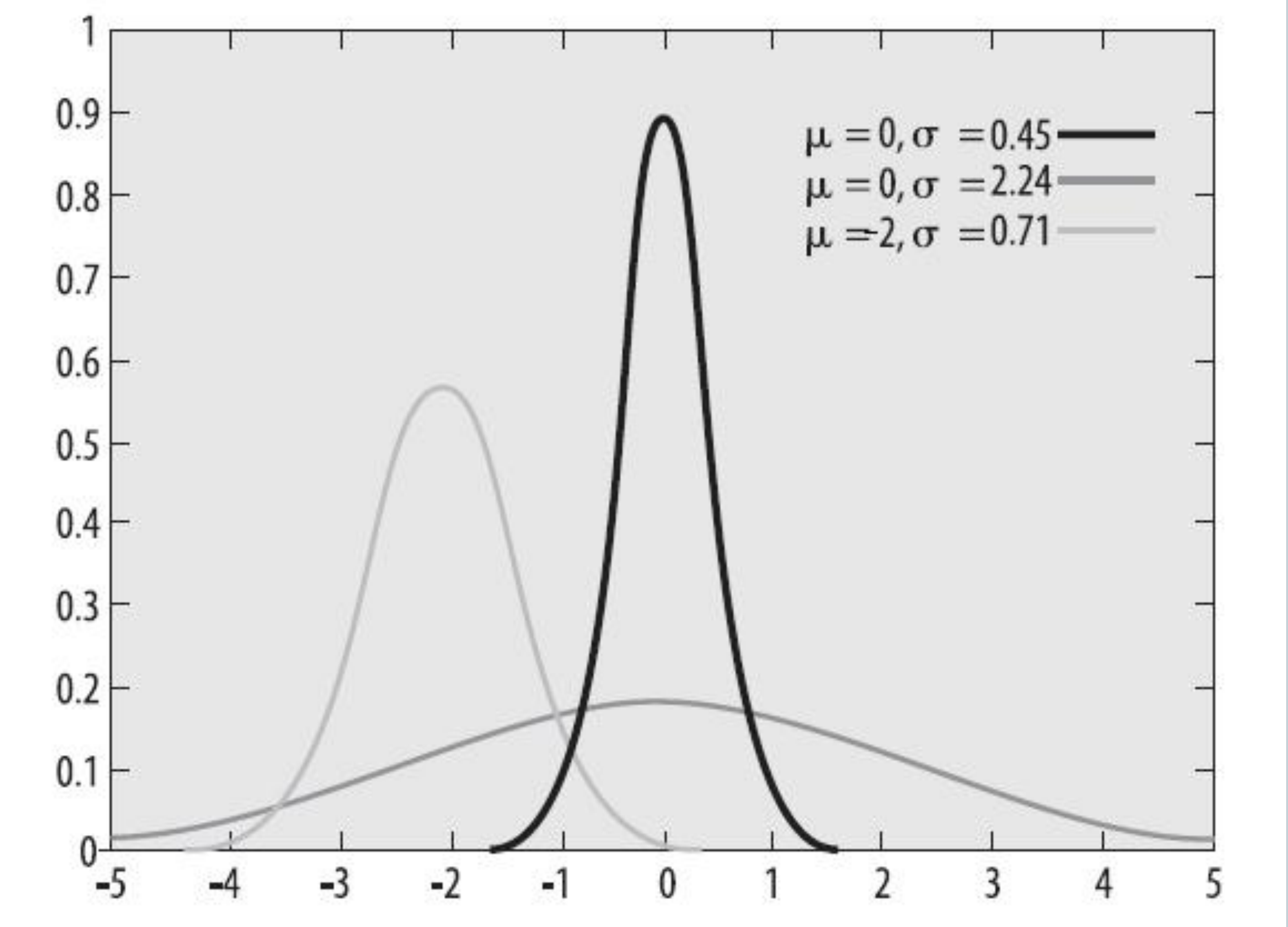
Boxplot and probability density function of a normal distribution $N(0, \sigma^2)$.

More probability density is found as one gets closer to the expected (mean) value in a normal distribution.

Normal Distribution

- The normal distribution is the most commonly used continuous distribution in statistics.
- There are an infinite number of normal distributions
  - They all have the same basic bell shape but differ according to their mean μ (the Greek letter *mu) and variance* σ (the Greek letter *sigma).*
- *Examples of three normal distributions with* different means and standard deviations are on the next slide

Three Normal Distributions



Normal Distribution

- The normal distribution with a mean of 0 and standard deviation of 1 is known as the *standard normal distribution or Z distribution*.
  - *Any normal distribution can be* transformed to the standard normal distribution by converting the original values to standardized scores.
- For example, weight measured in kilograms can be converted into Z scores,
  - express the value of the score in terms of units of the standard deviation

Normal Distribution

- Z-score formula: $Z = \frac{x - \mu}{\sigma}$
- If the variable *x is distributed normally with mean of 100 and standard deviation* of 5, i.e., *x ~ N(100, 5), a value of 105 has a Z-score of 1:*

$$Z = \frac{105 - 100}{5} = 1$$

- A value of 110 from this population has a Z-score of 2, and a value of 85 has a Z-score of –3.

Normal Distribution

- Z-scores facilitate comparison among populations
- with different means and standard deviations.
- Comparing our population *x ~ N(100, 5) with another population y ~ N(50,10), we can't* say whether a score of 95 among the first population is more or less unusual than a score of 35 among the second population.
- Using Z-scores: $Z = \dfrac{95-100}{5} = -1$ $Z = \dfrac{35-50}{10} = 1.5$
  - Standardized values can be used in multiple trait selection for animals, humans from different schools etc.

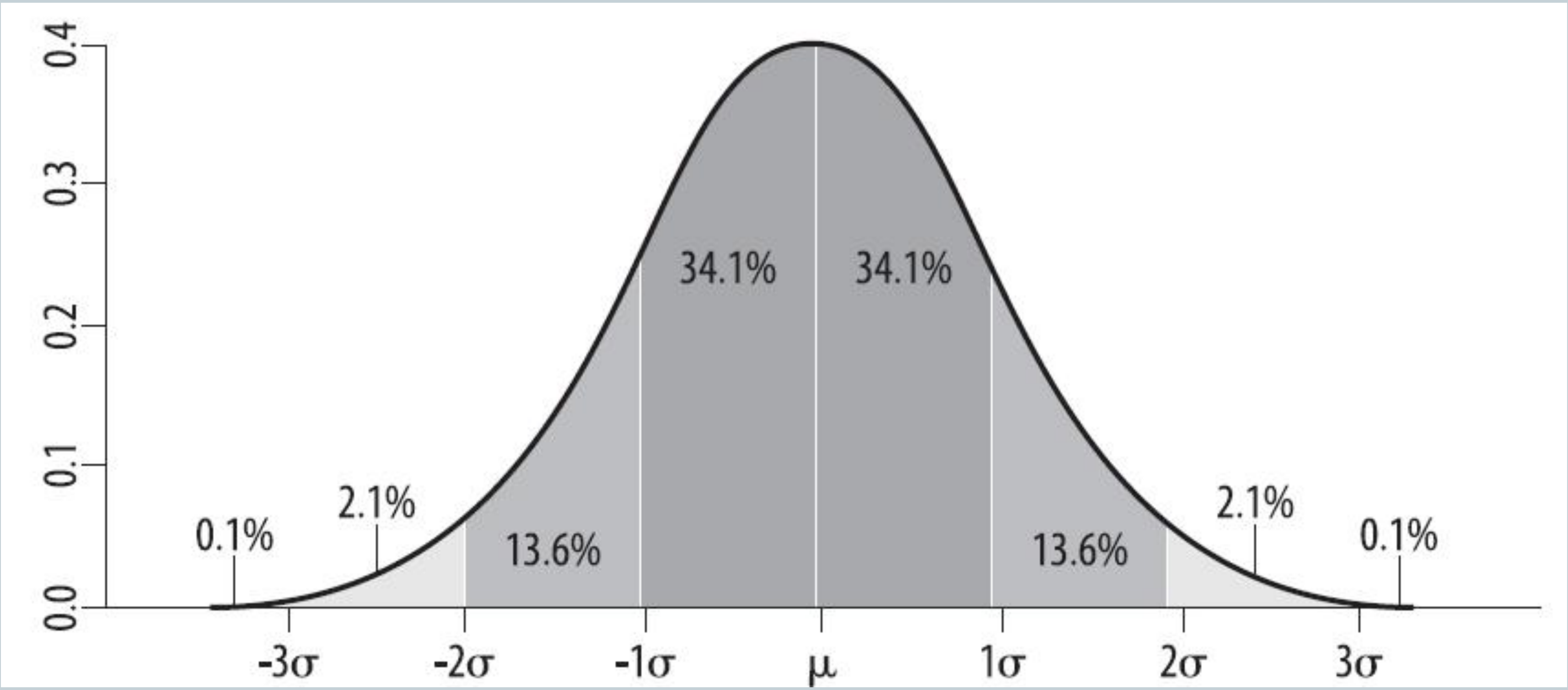
Normal Distribution

- The empirical rule states that for any normal distribution:
- • About 68% of the data will fall within one standard deviation of the mean
- • About 95% of the data will fall within two standard deviations of the mean
- • Over 99% of the data will fall within three standard deviations of the mean

Normal Distribution

- **Chebyshev's inequality:** in any probability distribution, "nearly all" values are close to the mean
- No more than $1/k^2$ of the distribution's values can be more than *k standard deviations* away from the mean
- = at least $1-1/k^2$ of the distribution's values are within *k* standard deviations of the mean.
- If k=2, at least 1-¼ = ¾ of the values within 2 std dev of the mean.
- If k=3, at least 1-1/9 = 8/9 are within 3 std dev.
- Only the case $k > 1$ is useful.
  - If k=1, at least 1-1=0 values are within 1 std ☺

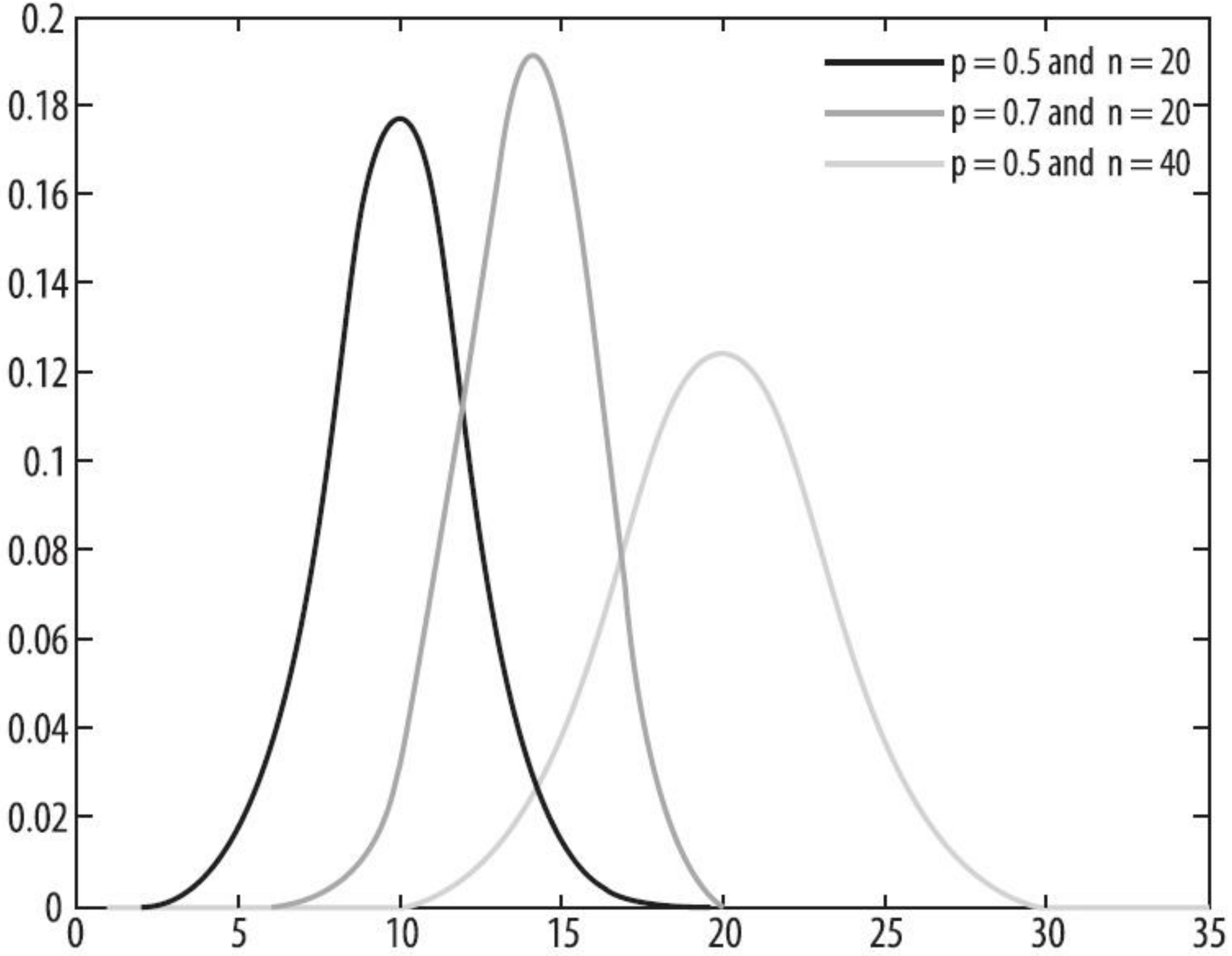
Specified ranges of the normal distribution



Binomial Distribution

- Most common discrete distribution
  - Male/female
  - Coin heads/Tails
- *k is the number of trials: if we are flipping a coin 10 times, $k = 10$.*
- *n is the number of successes: for instance, if we want to know the probability of 5* successes in 10 trials, *$n = 5$.*
- *p, a number between 0 and 1, is the probability of success*

Binomial Distribution



Binomial Distribution

- Common rule of thumb is that if both *np and n(1 – p) are greater* than 5, the binomial distribution may be approximated by the normal distribution.
- The distribution (*p = 0.5, n = 40) qualifies for the normal* approximation because:
- $np = 40(0.5) = 20$
- $n(1 - p) = 40(1 - 0.5) = 20$
- *Male and female may look like normal distribution if there are enough subjects.*

Independent and Dependent Variables

- Variables may be classified by the role they play in an experimental design.
- X variables, class variables, or independent variables
- Y variables, or dependent variables.
- Weight could be an independent variable, a covariate in one study, or a dependent variable in another one.

Independent and Dependent Variables

- In the first data set, total aerobic mesophilic bacteria count (TAC) was used as the dependent variable (Y) in the statistical analyses.
- Independent factors (X) were time, temperature and accelerating voltage (0 volts, 100 volts, 200 volts).

Independent and Dependent Variables

- In the second data set, yeast and mold, coliform bacteria, *E. coli* and *Staphylococcus aureus* counts were used as the dependent variables (Y).
- The treatment (initial load, control, AgNP100 and pasteurized milk) was the independent variable in the second data set.

Independent and Dependent Variables

The statistical model was:

- $Y_{ijklmno} = \mu + A_i + B_j + C_k + D_l + + F_o + e_{ijklmno}$

where;

- Y_{iklmno} = test day record for daily, or fat corrected daily milk yield, or percent fat content of the milk,
- μ = overall mean
- A_i = effects due to milking frequency (2x, 4x)
- B_j = effects due to parity (1, …, 5),
- C_k = effects due to birth type (single, multiple)
- D_l = effects due to test day (1, …, 7),
- X_m = covariates:
  - C_1 = DIM/c where c is constant, set to 300 days,
  - $C_2 = (DIM/c)^2$
  - C_3 = ln(c/DIM)
  - $C_4 = ((\ln(c/DIM))^2$, the subscript n denotes that regression were nested within parity
- F_o = effects of the subject,
- $e_{ijklmno}$ = random error.

Independent and Dependent Variables

- The *e means "error" and refers to the fact that we don't expect* any regression equation to perfectly predict *Y*.
- A lot of the statistics use greek letters to denote the variables in the equation.
- All the letters derives from this basic equation:
- $Y = \alpha + \beta X$
- Y=a+BX
- $Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \beta 3 X3 + \ldots + e$

Independent and Dependent Variables

- *In matrix form, the solution is:*
- *If* $Y = \beta X$
- $X'Y = \beta X'X$
- $\beta = (X'X)^{-1} (X'Y)$
- *Beta is the solution matrix*

Inferential Statistics

- In a randomized clinical trial of the effects of a new drug on hypertension
- if the correct procedures are followed and significant results are achieved
- the researcher can assert that changes in blood pressure observed were caused or influenced by the new drug.
- This assumes that all the other known factors were controlled carefully.
  - I had "researchers" with patients who smoked, but did not want to factor that into consideration.
  - Gives you false results, all that going into error.

The Central Limit Theorem

- The sampling distribution of the sample mean approximates the normal distribution, regardless of the distribution of the population from which the samples are drawn, <u>if the sample size is sufficiently large.</u>
    - Statistical inferences using tests based on the approximate normality of the mean, even if the sample is drawn from a population that is not normally distributed.

The Central Limit Theorem

- Let *X1,… Xn be a random sample from some population with mean μ and* variance σ2.
- Then for large *n,*

approximately distributed

$$\bar{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$$

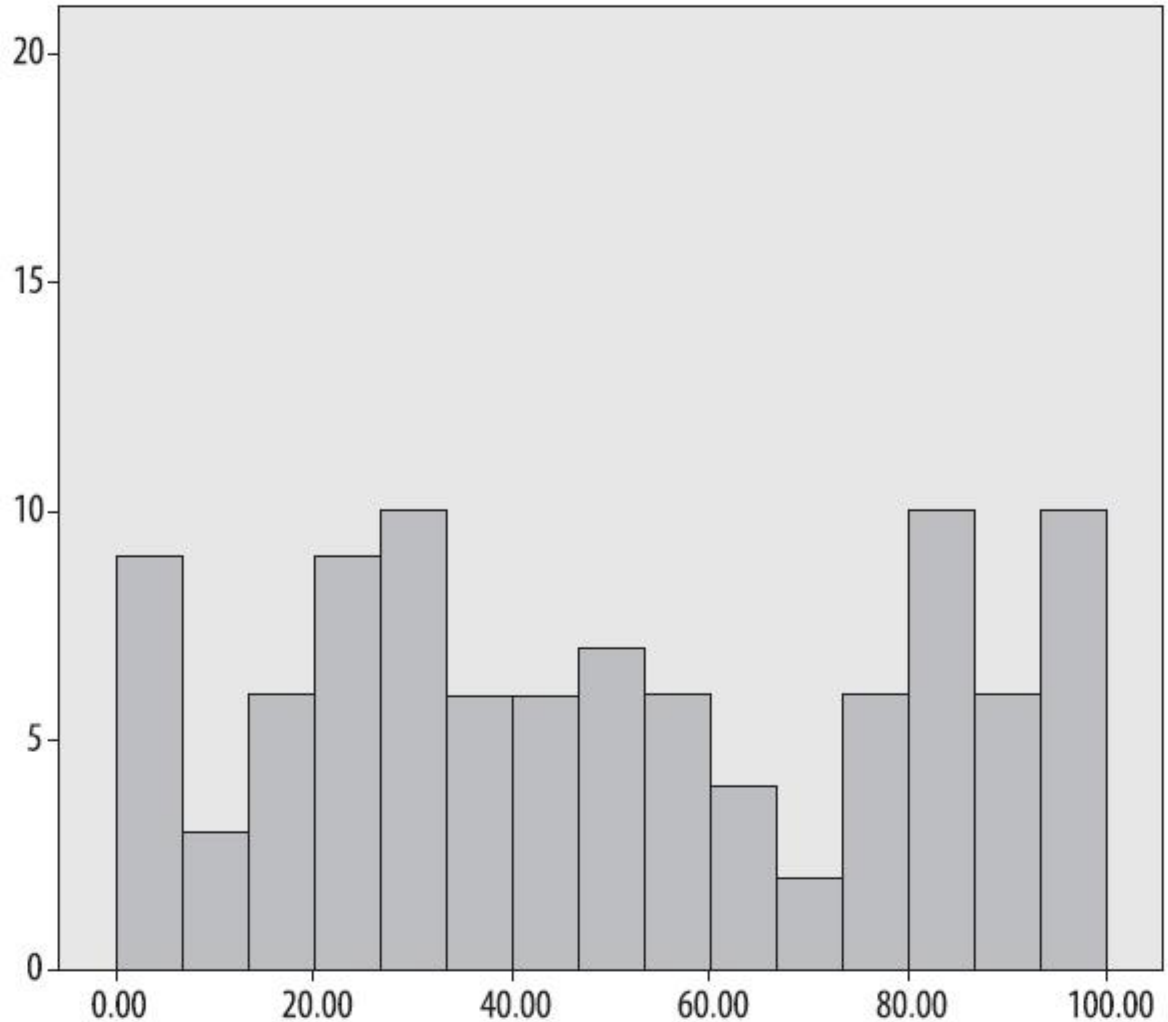
  - the mean of X is approximately normally distributed with mean μ and variance σ^2/n
- *Even if the distribution is not normal.*

The Central Limit Theorem

- The application of the Central Limit Theorem in practice can be seen through computer simulations that repeatedly draw samples of specified size from a nonnormal population.

Randomly generated data (100 cases) with a uniform distribution of values ranging from 0 to 100.

- If n=40
- then
- infinity ☺



Hypothesis Testing

- Hypothesis testing allows us to use statistical methods to make decisions about real-life problems:
  - Develop a research hypothesis that can be tested mathematically.
  - Formally state the null and alternative hypotheses.
  - Decide on an appropriate statistical test and do the calculations.
  - Make your decision based on the results.

Hypothesis Testing

- H_o: $\mu 1 = \mu 2$
- H_1: $\mu 1 \neq \mu 2$
- *Two tailed* alternative hypothesis:
- The blood pressure of patients treated with drug A is different, either higher or lower, than that of patients receiving standard treatment.
- Reject the null hypothesis and accept the alternative hypothesis (Significant results)
- Fail to reject the null hypothesis ($P<0.05$)
  - Failure to reject the null hypothesis does not mean that we have proven it to be true, only that the experiment or study did not find sufficient evidence to reject it.

Hypothesis Testing

- The P-value is commonly set at 0.05. Why this particular value?
- It's an arbitrary cutoff point and dates back to the early twentieth century, when statistics were computed by hand and the results compared to published tables used to determine whether a result was significant or not.
- The use of $P < 0.05$ still remains common practice for
- published research.
- Alternative lower values are sometimes used, such as $P < 0.01$ or $P < 0.001$, but no one has been successful in legitimizing the use of a higher cutoff, such as $P < 0.10$.

Hypothesis Testing

- A lot of drug trials only reports significant results, while non significant results are buried under a rock, because journals tend to accept only significant results.
- If you have a drug that has no effect on something, beneficial or harmful, it is still a result because then other people won't have to try the same thing,
- But it is considered "boring" by a lot of evaluators.
- If you have a large enough sample, any little difference will be statistically significant.

Type I and Type II error

Reality	H_o Accept	H_o Reject
H_o accept	Probability of correct decision=1-α	Probability of Type I error = α
H_o reject	Probability of Type II error = β	Probability of correct decision=1-β (Power)

- Type I error is detecting an effect that is not present
- Setting $\alpha = 0.05$ = 5% chance of rejecting the null hypothesis when we should have accepted it.
- Type II error is failing to detect an effect that is present.

Type I and Type II error

- Type II error is considered a less serious error.
  - You find no differences, and state that we failed to reject the null hypothesis as opposed to finding differences that are not there and stating that we reject the null hypothesis.
  - Failing to make an inference that is true (Type II error) compared to making an inference that is false (Type I error).
- Conventional levels of acceptability for Type II error are $\beta = 0.1$ or $\beta = 0.2$.
- If $\beta = 0.1$, the study has a 10% probability of a Type II error
  - There is a 10% chance that the null hypothesis will be false but will be accepted in the study.

Type I and Type II error

- The reverse of Type II error is *power, defined as* $1 - \beta$.
- *The importance of* setting an appropriate power level has become more appreciated in recent years, especially in the medical field.
- Researchers, ethic boards and funding agencies have become concerned with power, and thus with Type II error
  - They don't want to spend the time and effort (and animals) to conduct a study unless it has a reasonable probability of finding existing significant results.

Confidence Intervals

- The sample mean is a point estimate
- The point estimate is likely to vary by chance if we had chosen a different sample
- Many people report both point estimates and interval estimates.
- A point estimate is a single number, while an interval estimate is a range or interval of numbers.

Confidence Intervals

- The confidence interval is calculated using α
- The *confidence coefficient is calculated as (1 – α) or, as a* percentage, 100(1 – α)%.
- Thus if $\alpha = 0.05$, the confidence coefficient is 0.95 or 95%.
- The latter usage is more common; for instance, professional journals often require that you report the 95% confidence interval for your statistics.

Confidence Intervals

- The confidence interval conveys important information about the precision of the point estimate.
- Two samples of students and in both cases the mean IQ score is 100.
- 95% confidence interval in group one is (95, 105),
- 95% confidence interval in the other is (80, 120).
- Because the former confidence interval is much narrower than the latter, the estimate of the mean is more precise for the first sample.

P value

- We use to “approximate” P values based on the table values of pre-calculated P values.
- Sampling error is a possibility in studies based on samples.
- We want to know the probability that the results obtained from our sample were not due to chance.
  - If we had the means to draw repeated samples from the population and repeat the experiment, how likely is it that we would obtain similar results most of the time?

P value

- *P-value is the probability that results at least as extreme as those* obtained in a sample were due to chance.
- At least as extreme is there because most statistical tests involve comparing the test statistic to, for example the normal distribution
  - Scores closer to the center of the distribution are most common and scores become less likely as they are further from the center of the distribution.

P value

- We flip a fair coin 10 times and 8 times it comes up heads.
- We want to know the *p-value of this result*
  - *How* likely is it that a coin with a probability of 0.5 for heads on any single trial would produce 8 heads in 10 trials?

P value

- Using a binomial table, computer software, or the binomial formula, we find that the probability of this exact result (8 heads in 10 trials) is 0.0439
  - less than 5% of the time would we expect to get exactly 8 heads in 10 flips with a fair coin.
- The probability for 9 heads in 10 trials is 0.0098
- The probability for 10 heads in 10 trials is 0.0010.
- This demonstrates that as results move further away from the expected result of 5 heads in 10 trials, they become less likely.
- Or the differences are large enough to be significant.

P value

- A fair coin= heads or tails are equally likely outcomes for any single flip
- Evaluate the probability that the coin truly is fair
  - Results that are far from our expectation give us strong evidence that it in fact is not fair.
- We calculate the probability not just of the result we obtained in our experiment, but of results at least as extreme as those we obtained.
- In this case, the probability of getting 8, 9 or 10 heads in 10 flips of a fair coin is 0.0439 + 0.0098 + 0.0010= 0.0547 (5 percent).
  - A one in 20 value, a significant value: it is far from expected
  - P-value for the result of at least 8 heads in 10 trials using a coin.

P value

- *P-values are commonly reported for most research*
  - intuition is a poor guide to how unusual a particular result is.
- Many people might think it is unusual to get 8 or more heads on 10 trials using a fair coin.
- In this case, the binomial probability of such a result has a *p-value of 0.0547*.
- *This result does not allow us to reject the null* hypothesis that the coin is fair, i.e., *P(heads) = 0.5, using the standard rule of* thumb that a *p-value must be less than 0.05 for results to be considered* significant.

P value

- If we calculated 9 or 10 heads, it would come out
  - Significant
  - Highly unusual
  - Not small enough to be attributed to chance
  - Large enough that it cannot be attributed to chance.
- The probability of getting 9 or 10 heads in 10 flips of a fair coin is 0.0098 + 0.0010= 0.01