# Statistical Significance and Univariate and Bivariate Tests

BUS 735: Business Decision Making and Research

# 1

## 1.1 Goals

### Goals

- Specific goals:
  - Re-familiarize ourselves with basic statistics ideas: sampling distributions, hypothesis tests, p-values.
  - Be able to distinguish different types of data and prescribe appropriate statistical methods.
  - Conduct a number of hypothesis tests using methods appropriate for questions involving only one or two variables.
- Learning objectives:
  - LO1: Be able to construct and test hypotheses using a variety of bivariate statistical methods to compare characteristics between two populations.
  - LO6: Be able to use standard computer packages such as SPSS and Excel to conduct the quantitative analyses described in the learning objectives above.

# 2 Statistical Significance

## 2.1 Sampling Distribution

### Probability Distribution

- **Probability distribution:** summary of all possible values a variable can take along with the probabilities in which they occur.
- Usually displayed as:

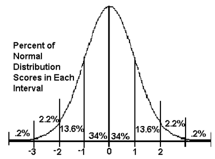
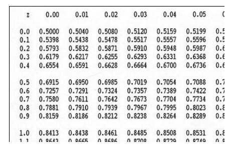
**Picture** 


**Table** 

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 |

**Formula**

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- **Normal distribution:** often used "bell shaped curve", reveals probabilities based on how many standard deviations away an event is from the mean.

### Sampling distribution

- Imagine taking a sample of size 100 from a population and computing some kind of statistic.
- The statistic you compute can be anything, such as: mean, median, proportion, difference between two sample means, standard deviation, variance, or anything else you might imagine.
- Suppose you repeated this experiment over and over: take a sample of 100 and compute and record the statistic.
- A **sampling distribution** is the probability distribution *of the statistic*
- Is this the same thing as the probability distribution of the population? NO! They may coincidentally have the same shape though.

### Example

- Sampling Distribution Simulator
- In reality, you only do an experiment once, so the sampling distribution is a hypothetical distribution.
- Why are we interested in this?

**Desirable qualities**

What are some qualities you would like to see in a sampling distribution?

- The average of the sample statistics is equal to the true population parameter.
- Want the variance *of the sampling distribution* to be as small as possible. Why?
- Want the *sampling distribution* to be normal, regardless of the distribution of the population.

## 2.2 Central Limit Theorem

**Central Limit Theorem**

- Given:
  - Suppose a RV $x$ has a distribution (it need not be normal) with mean $\mu$ and standard deviation $\sigma$.
  - Suppose a *sample mean* ($\bar{x}$) is computed from a sample of size $n$.
- Then, if $n$ is sufficiently large, the sampling distribution of $\bar{x}$ will have the following properties:
  - The sampling distribution of $\bar{x}$ will be normal.
  - The mean of the sampling distribution will equal the mean of the population (consistent):
    $$\mu_{\bar{x}} = \mu$$
  - The standard deviation of the sampling distribution will decrease with larger sample sizes, and is given by:
    $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

**Central Limit Theorem: Small samples**

If $n$ is small (rule of thumb for a single variable: $n < 30$)

- The sample mean is still consistent.
- Sampling distribution will be normal if the distribution of the population is normal.

**Example 1**

Suppose average birth weight is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.

What is the probability that a sample of size $n = 30$ will have a mean of $7.5lbs$ or greater?

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{7.5 - 7}{1.5/\sqrt{30}} = 1.826$$

The probability the sample mean is greater than $7.5lbs$ is:

$$P(\bar{x} > 7.5) = P(z > 1.826) = 0.0339$$

**Example 2**

Suppose average birth weight is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.

What is the probability that a randomly selected baby will have a weight of 7.5lbs or more? What do you need to assume to answer this question?
Must assume the population is normally distributed. Why?

$$z = \frac{x - \mu}{\sigma} = \frac{7.5 - 7}{1.5} = 0.33$$

The probability that a baby is greater than 7.5lbs is:

$$P(x > 7.5) = P(z > 0.33) = 0.3707$$

**Example 3**

- Suppose average birth weight of all babies is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.
- Suppose you collect a sample of 30 newborn babies whose mothers used illegal drugs during pregnancy.
- Suppose you obtained a sample mean $\bar{x} = 6lbs$. If you assume the mean birth weight of babies whose mothers used illegal drugs has the same sampling distribution as the rest of the population, what is the probability of getting a sample mean this low or lower?

**Example 3 continued**

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$
$$z = \frac{6 - 7}{1.5/\sqrt{30}} = -3.65$$

The probability the sample mean is less than or equal to $6lbs$ is:

$$P(\bar{x} < 6) = P(z < -3.65) = 0.000131$$

That is, if using drugs during pregnancy actually does still lead to an average birth weight of 7 pounds, there was only a 0.000131 (or 0.0131%) chance of getting a sample mean as low as six or lower. This is an extremely unlikely event if the assumption is true. Therefore it is likely the assumption is not true.

## 2.3 Hypotheses Tests

**Statistical Hypotheses**

- A **hypothesis** is a claim or statement about a property of a population.
  - Example: The population mean for systolic blood pressure is 120.
- A **hypothesis test** (or **test of significance**) is a standard procedure for testing a claim about a property of a population.
- Recall the example about birth weights with mothers who use drugs.
  - Hypothesis: Using drugs during pregnancy leads to an average birth weight of 7 pounds (the same as with mothers who do not use drugs).

**Null and Alternative Hypotheses**

- The **null hypothesis** is a statement that the value of a population parameter (such as the population mean) *is equal to* some claimed value.
  - $H_0$: $\mu = 7$.
- The **alternative hypothesis** is an alternative to the null hypothesis; a statement that says a parameter differs from the value given in the null hypothesis.
  - $H_a$: $\mu < 7$.
  - $H_a$: $\mu > 7$.
  - $H_a$: $\mu \neq 7$.
- In hypothesis testing, assume the null hypothesis is true until there is strong statistical evidence to suggest the alternative hypothesis.
- Similar to an "innocent until proven guilty" policy.

**Hypothesis tests**

- (Many) hypothesis tests are all the same:

$$z \text{ or } t = \frac{\text{sample statistic} - \text{null hypothesis value}}{\text{standard deviation of the sampling distribution}}$$

- Example: hypothesis testing about $\mu$:
  - Sample statistic = $\bar{x}$.
  - Standard deviation of the sampling distribution of $\bar{x}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

**P-values**

- The P-value of the test statistic, is the area of the sampling distribution from the sample result in the direction of the alternative hypothesis.
- Interpretation: If the null hypothesis is correct, than the p-value is the probability of obtaining a sample that yielded your statistic, or a statistic that provides even stronger evidence of the null hypothesis.
- The p-value is therefore a measure of *statistical significance.*
  - If p-values are very small, there is strong statistical evidence in favor of the alternative hypothesis.
  - If p-values are large, there is insignificant statistical evidence. When large, you fail to reject the null hypothesis.
- Best practice is writing research: report the p-value. Different readers may have different opinions about how small a p-value should be before saying your results are statistically significant.

# 3 Univariate Tests

## 3.1 Types of Data/Tests

**Types of Data**

- Nominal data: consists of categories that cannot be ordered in a meaningful way.
- Ordinal data: order is meaningful, but not the distances between data values.
  - Excellent, Very good, Good, Poor, Very poor.

- Interval data: order is meaningful, *and* distances are meaningful. However, there is *no natural zero*.
  - Examples: temperature, time.
- Ratio data: order, differences, and zero are all meaningful.
  - Examples: weight, prices, speed.

**Types of Tests**

- Different types of data require different statistical methods.
- Why? With interval data and below, operations like addition, subtraction, multiplication, and division are *meaningless!*
- Parametric statistics:
  - Typically take advantage of central limit theorem (imposes requirements on probability distributions)
  - Appropriate only for interval and ratio data.
  - More **powerful** than nonparametric methods.
- Nonparametric statistics:
  - Do not require assumptions concerning the probability distribution for the population.
  - There are many methods appropriate for ordinal data, some methods appropriate for nominal data.
  - Computations typically make use of data's *ranks* instead of actual data.

## 3.2 Hypothesis Testing about Mean

**Single Mean T-Test**

- Test whether the population mean is equal or different to some value.
- Uses the sample mean its statistic.
- Why T-test instead of Z-test?
- Parametric test that depends on results from Central Limit Theorem.
- Hypotheses
  - Null: The population mean is equal to some specified value.
  - Alternative: The population mean is [greater/less/different] than the value in the null.

**Example: Public School Spending**

- Dataset: average pay for public school teachers and average public school spending per pupil for each state and the District of Columbia in 1985.
- Download dataset `eduspending.xls`.
- Conduct the following exercises:
  - Show some descriptive statistics for teacher pay and expenditure per pupil.
  - Is there statistical evidence that teachers make less than $25,000 per year?
  - Is there statistical evidence that expenditure per pupil is more than $3,500?

## 3.3 Hypothesis Testing about Proportion

**Single Proportion T-Test**

- **Proportion:** Percentage of times some characteristic occurs.
- Example: percentage of consumers of soda who prefer Pepsi over Coke.

$$\text{Sample proportion} = \frac{\text{Number of items that has characteristic}}{\text{sample size}}$$

**Example: Economic Outlook**

- Data from Montana residents in 1992 concerning their outlook for the economy.
- All data is ordinal or nominal:
  - AGE = 1 under 35, 2 35-54, 3 55 and over
  - SEX = 0 male, 1 female
  - INC = yearly income: 1 under $20K, 2 20-35$K, 3 over $35K
  - POL = 1 Democrat, 2 Independent, 3 Republican
  - AREA = 1 Western, 2 Northeastern, 3 Southeastern Montana
  - FIN = Financial status 1 worse, 2 same, 3 better than a year ago
  - STAT = 0, State economic outlook better, 1 not better than a year ago
- Do the majority of Montana residents feel their financial status is the same or better than one year ago?
- Do the majority of Montana residents have a more positive economic outlook than one year ago?

## 3.4 Nonparametric Testing about Median

**Single Median Nonparametric Test**

- Why?
  - Ordinal data: cannot compute sample means (they are meaningless), only median is meaningful.
  - Small sample size and you are not sure the population is not normal.
- Sign test: can use tests for proportions for testing the median.
  - For a null hypothesized population median...
  - Count how many observations are above the median.
  - Test whether that proportion is greater, less than, or not equal to 0.5.
  - For small sample sizes, use binomial distribution instead of normal distribution.

**Example: Attitudes Grade School Kids**

- Dataset: 438 students in grades 4 through 6 were sampled from three school districts in Michigan. Students ranked from 1 (most important) to 5 (least important) how important grades, sports, being good looking, and having lots of money were to each of them.
- Open dataset `gradschools.xls`. Choose second worksheet, titled `Data`.
- Answer some of these questions:
  - Is the median importance for grades is greater than 3?
  - Is the median importance for money less than 3?

# 4 Bivariate Tests

## 4.1 Correlation

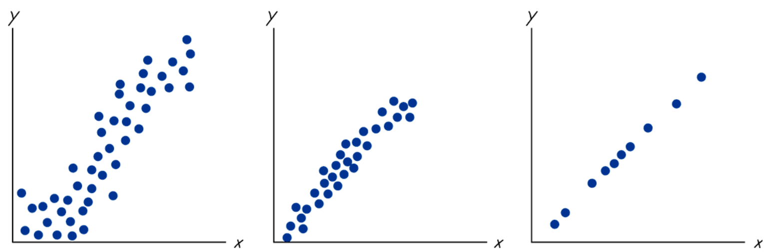
**Correlation**

- A **correlation** exists between two variables when one of them is related to the other in some way.
- The **Pearson linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.
  - Parametric test!
  - Null hypothesis: there is zero linear correlation between two variables.

  - Alternative hypothesis: there is [positive/negative/either] correlation between two variables.

- Spearman's Rank Test
  - Non-parametric test.
  - Behind the scenes - replaces actual data with their *rank*, computes the Pearson using ranks.
  - Same hypotheses.

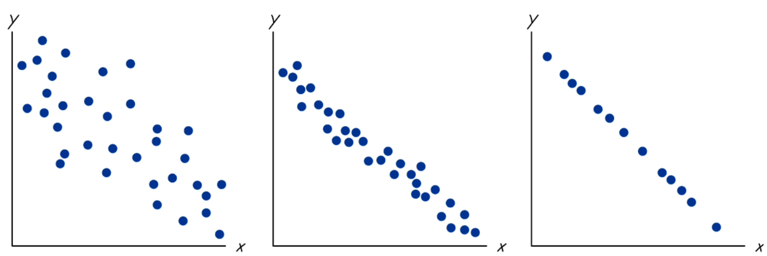
### Positive linear correlation



**(a)** Positive correlation between $x$ and $y$ **(b)** Strong positive correlation between $x$ and $y$ **(c)** Perfect positive correlation between $x$ and $y$

- Positive correlation: two variables move in the same direction.
- Stronger the correlation: closer the correlation coefficient is to 1.
- Perfect positive correlation: $\rho = 1$

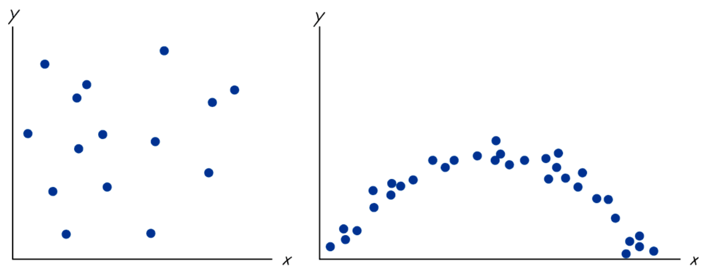
### Negative linear correlation



**(d)** Negative correlation between $x$ and $y$ **(e)** Strong negative correlation between $x$ and $y$ **(f)** Perfect negative correlation between $x$ and $y$

- Negative correlation: two variables move in opposite directions.
- Stronger the correlation: closer the correlation coefficient is to -1.
- Perfect negative correlation: $\rho = -1$

**No linear correlation**




**(g)** No correlation between $x$ and $y$

**(h)** Nonlinear relationship between $x$ and $y$

- Panel (g): no relationship at all.
- Panel (h): strong relationship, but not a *linear* relationship.
  - Cannot use regular correlation to detect this.

**Example: Public Expenditure**

- Data from 1960! about public expenditures per capita, and variables that may influence it:
  - Economic Ability Index
  - Percentage of people living in metropolitan areas.
  - Percentage growth rate of population from 1950-1960.
  - Percentage of population between the ages of 5-19.
  - Percentage of population over the age of 65.
  - Dummy variable: Western state (1) or not (0).
- Is there a statistically significant linear correlation between the percentage of the population who is young and the public expenditure per capita?
- Is there a statistically significant linear correlation between the public expenditure per capita and whether or not the state is a western state?

## 4.2 Difference in Populations (Independent Samples)

**Difference in Means (Independent Samples)**

- Suppose you want to know whether the mean from one population is larger than the mean for another.
- Examples:
  - Compare sales volume for stores that advertise versus those that do not.

  - Compare production volume for employees that have completed some type of training versus those who have not.

- Statistic: Difference in the sample means ($\bar{x}_1 - \bar{x}_2$).

**Independent Samples T-Test**

- Hypotheses:
  - Null hypothesis: the difference between the two means is zero.
  - Alternative hypothesis: the difference is [above/below/not equal] to zero.
- Different ways to compute the test depending on whether...
  - the variance in the two populations is the same (more powerful test), or...
  - the variance of the two populations is different.
  - To guide you, SPSS also reports Levene's test for equality of variance (Null - variances are the same).

**Example**

- Dataset: average pay for public school teachers and average public school spending per pupil for each state and the District of Columbia in 1985.
- Test the following hypotheses:
  - Does spending per pupil differ in the North (region 1) and the South (region 2)?
  - Does teacher salary differ in the North and the West (region 3)?
- Do you see any weaknesses in our statistical analysis?

**Nonparametric Tests for Differences in Medians**

- Mann-Whitney U test: nonparametric test to determine difference in *medians*.
- Assumptions:
  - Samples are independent of one another.
  - The underlying distributions have the same shape (i.e. only the location of the distribution is different).
  - It has been argued that violating the second assumption does not severely change the sampling distribution of the Mann-Whitney U test.
- Null hypothesis: medians for the two populations are the same.
- Alternative hypotheses: medians for the two populations are different.

## 4.3 Paired Samples

### Dependent Samples - Paired Samples

- Use a **paired sampled test** if instead the two samples have the same individuals before and after some treatment.
- Really simple: for each individual subtract the before treatment measure from the after treatment measure (or vice-versa).
- Treat your new series as a single series.
- Conduct one-sample tests.
- In SPSS, you need to have separate columns for each of these variables.
- There are methods in SPSS specifically for Dependent Samples tests - but the paired sampled approaches are equivalent to one-sample tests.

# 5

### Conclusions

- Ideas to keep in mind:
  - What is a sampling distribution? What does it imply about p-values and statistical significance?
  - When it is appropriate to use parametric versus non-parametric methods.
  - Most univariate and bivariate questions have a parametric and non-parametric approach.
- Homework posted on the class website.
- Next: Regression Analysis - looking at more complex relationships between more than 2 variables.