Statistics Review

Professor Song Yao
Olin Business School

Customer Analytics

Describe Data vs. Inference

We will use statistics for two types of tasks

STATISTICAL TASKS (with examples)

- **Describe data (a sample of the customers)**
  - "How much do our customers spend each month on average?"
  - "What percentage of our customers are unprofitable?"
  - "What is the difference between the response rates of men and women?"
- **Inference (draw conclusions about all customers) from data (sample)**
  - "Based on our sample, does the difference between the response rates of men and women indicate that men and women respond differently in the customer base at large?"
  - "Based on our test mailing, can we conclude that ad-copy A works more effectively than ad-copy B?"

It is useful to distinguish between different types of variables

VARIABLE DEFINITIONS

- Assumes different values across population members
  - Ex: age, salary, ethnicity
- **Non-metric (Non-numeric)**
  - Categorical or nominal (gender, zip code, brand, religion)
  - Ordinal (Business Week rankings, NCAA rankings)
- **Metric (Numeric)**
  - Sales, price, age, rainfall
- Different types of variables are handled in different ways in statistics
  - Can talk about an average age, but not an average color
  - Some statistical techniques only work with one type of variable

The most simple descriptive statistics summarize one variable only

BASIC DESCRIPTIVE STATISTICS

– **Measures of "central tendency"**
  - Numeric data: Mean, median
  - Non-numeric data: mode

– **Measures of dispersion**
  - Numeric data: Variance, standard deviation, range

– **Distribution/shape of the data** (histograms and bar charts)

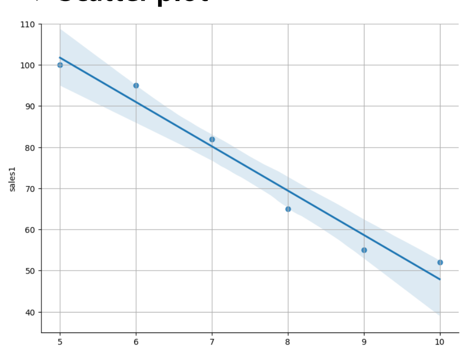
– Aside: ignoring 1 or 2 of the above is a common way to 'lie' with statistics

We are often interested in describing how two variables are related / associated with each other

EXAMPLES OF VARIABLE ASSOCIATION

"In our test market, are higher prices associated with lower sales?"

--> Scatter plot



"In our e-mail campaign, what is the difference between the responses of men and women?"

--> **Cross Tabulation (cross-tab)**

Gender	Responded? yes	no	Total
Male	320	5020	5400
Female	384	4216	4600
Total	704	9216	10000

"In our test mailing, do urban customers spend more than rural customers?"

--> Mean comparison

Location	$ Order (mean)
Urban	35.10
Rural	31.34
Diff.	3.76

Most of the time, we want to draw conclusions from our sample about the population at large

INFERENCE PROBLEM IN STATISTICS

- Know that **in the test mailing** urban customers spent more than rural customers, \$35.10 vs. \$31.34
- Can we conclude from this that we can reliably expect that this difference exists **in our customer base at large?**
  (Or is this a "fluke" of our test mailing?)

=

"Is the average spending of rural customers
statistically significantly different
from the average spending of urban customers?"

Statistical inference allows us to make conclusions about the population at large
Note: *Statistical inference is different from causal inference*

Inference through Hypothesis Testing

Statistical inference is to decide which one of two hypotheses (scenarios) is more likely to be true

TWO HYPOTHESES

Hypothesis 0

– Average expenditures of urban and rural customers are **the same**

If this is true then we'd expect that…

$AvgExp_{urban} = AvgExp_{rural}$
(or at least close to equal)

$AvgExp_{urban} - AvgExp_{rural} = 0$
(or at least close to 0)

Hypothesis 1

– Average expenditures of urban and rural customers are **different**

If this is true then we'd expect that…

$AvgExp_{urban} \neq AvgExp_{rural}$

$AvgExp_{urban} - AvgExp_{rural} \neq 0$

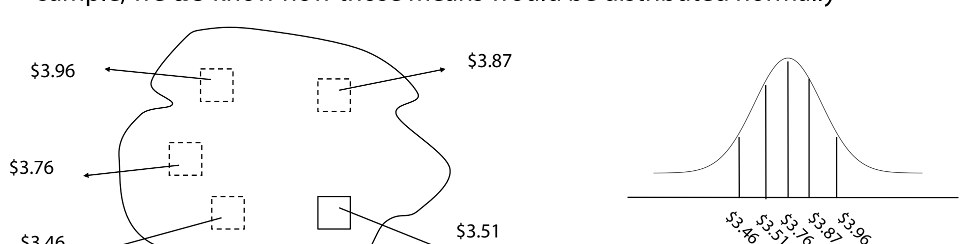
In our example, $AvgExp_{urban} - AvgExp_{rural}$ = \$3.76. Problem: Is this close to 0 enough?

Solution: "Statistics" enables us to conclude whether a value is "close to" or "far from" zero

STATISTICS AND INFERENCE

– Technically, a "statistic" is a number derived from a formula based on data
  - A mean is a statistic: the sum of the variable values divided by the number of observations
  - A variance is a statistic

– Amazingly, we know things about statistics for certain (those formula-based numbers based on data), independent of the underlying population itself
  - For example, we might not know how the expenditures of the customers in the urban customer base are distributed
  - But, if we were to take 5 sets of random samples, and calculated the mean of each sample, we **do** know how those means would be distributed normally



How does knowing the distribution of a statistic help us?

HOW A STATISTICAL HYPOTHESIS TEST WORKS

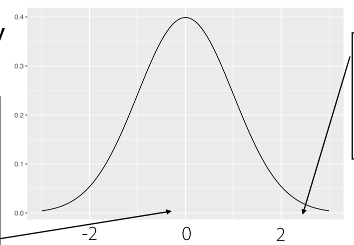
For a given random sample

- We can compute the difference of mean spendings ($AvgExp_{urban}$ - $AvgExp_{rural}$)
- We can further compute a ***t-statistics*** based on ($AvgExp_{urban}$ - $AvgExp_{rural}$)

We also know that

- **If** the average expenditures of urban and rural customers is the same (if Hypothesis 0 is true)
- **Then** t-statistic have a particular distribution that is centered around zero

Probability



t-statistic

If the value of the t-statistic for our sample falls here, then we know this is a pretty *likely* outcome when Hypothesis 0 is true, so we accept Hypothesis 0.

If the value of the t-statistic for our sample falls here, then we know it is a pretty *unlikely* outcome if Hypothesis 0 is true. So we reject Hypothesis 0

The power of statistical tools is that we know many statistics, so we can test many kinds of hypotheses

LOGIC OF ALL STATISTICAL TESTS

- If Hypothesis 0 is true, then a certain range of values of a particular statistic (formula-based numbers computed from sample data) is likely to occur and another range of values is unlikely to occur
- Calculate the "test-statistic" (or let R/Python do it for you)
- Check whether the value of the test-statistic falls in the "likely" or "unlikely" range (or let R/Python do it for you)
- If the test-statistic is in the "likely" range, we conclude Hypothesis 0 is probably true.
- If it is in the "unlikely" range, we conclude Hypothesis 0 is probably not true.
- The "unlikely" range is usually defined as a **less than 5% chance of observing the test-statistic in that range**
- Whether the chance is less than 5% or not is typically expressed as a "p-value"
  - p-value≥0.05 --> test-statistic in likely range --> "accept" Hypothesis 0
  - p-value<0.05 --> test-statistic in unlikely range --> reject Hypothesis 0

Confidence Interval

Prelude to Confidence Interval: A Numerical Example

```
# There are about 300,000 high-school students in Missouri, whose ages are
# (almost) uniformly distributed between 14 and 19 (NOT normal distribution!).
```

Out of the 300K students, we (1) randomly select 100 students, (2) record their ages, and (3) compute the 100 students' average age. We repeat this process 1,000 times, resulting in 1,000 average ages of 1,000 samples.

	sample_1	sample_2	sample_3	sample_4	sample_5	sample_6
0	15.18	14.56	16.56	14.49	15.19	18.39
1	16.71	15.92	16.77	14.17	16.12	18.77
2	15.80	17.49	17.66	15.20	14.84	14.10
3	14.12	18.22	17.96	14.03	15.88	14.70
4	16.48	15.40	14.37	14.79	14.50	14.67

… (1K columns)

… (100 rows)

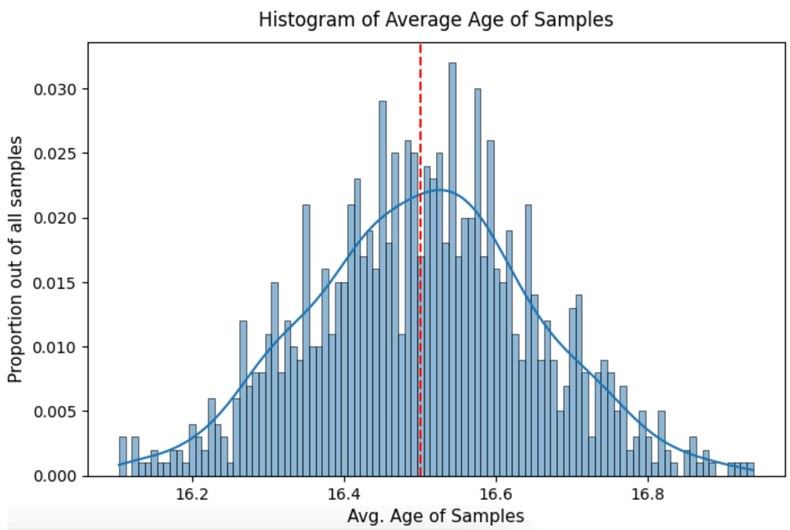
Here is the distribution of the average ages of the 1K samples

```
# Draw a histogram of the average ages of the samples

# first compute the average age of each sample
avg_age_dat = age_samples.mean(axis=0, skipna=True).reset_index()
avg_age_dat = avg_age_dat.rename(columns={"index": "sample", 0: "average_age"})

plt.figure(figsize=(8, 5))
sns.histplot(data=avg_age_dat, x="average_age", bins=100, stat="probability", kde=True)
plt.xlabel("Avg. Age of Samples", size=11)
plt.ylabel("Proportion out of all samples", size=11)
plt.title("Histogram of Average Age of Samples", pad=10)

# add a vertical line to show the population average age, 17
plt.axvline(x=16.5, color='red', linestyle='--', label="Population average age")
plt.show()
```



The graph has illustrated one of the most fundamental results in statistics. But first, concept review!

REVIEW OF CONCEPTS BEFORE KEY RESULTS

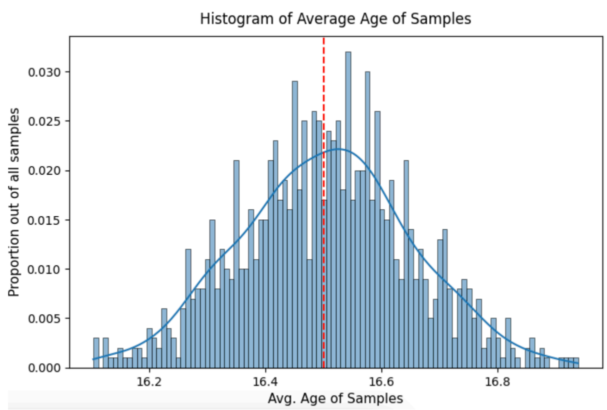
- What's the difference between ***the standard deviation of a sample*** and ***the standard error of sample mean***?

- Simply put
  - ***The standard deviation of a sample*** is about sample "dispersion," the degree to which observations deviate from the ***sample average*** (SA).
  - ***The standard error of sample mean*** is an estimate of how far the ***sample average*** is away from the ***true average*** (TA) of the population.

The graph has illustrated one of the most fundamental results in statistics

KEY RESULTS

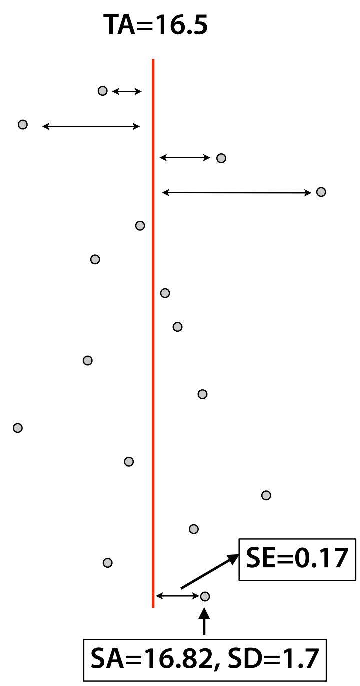
- When drawing random samples from a population, the averages of the samples will follow the normal curve (if the sample is reasonably large)
- The average difference between the true average (TA) and sample averages (SA) (the **"standard error"**) is :
  - SE = Standard Deviation / Sqrt(Sample size)



Consider the following example

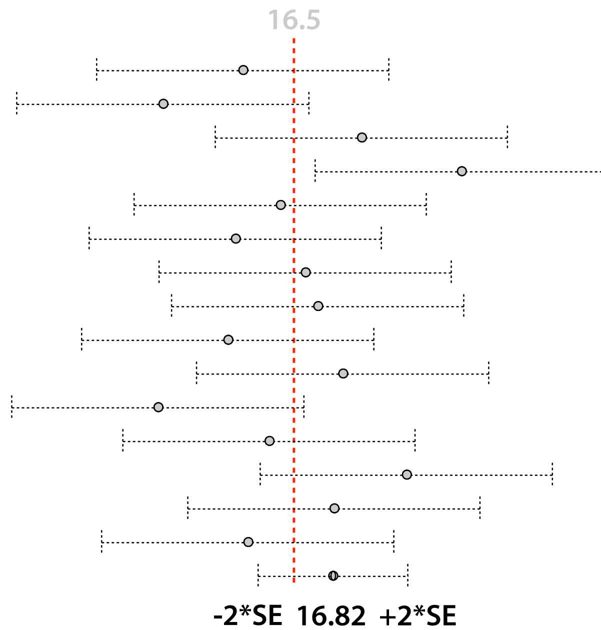
EXAMPLE: In the simulation we just saw

- True average=16.5
- One of the samples (the last one) has a mean of 16.82 and SD=1.7
- The sample size is of 100
- Then the average difference between SA and TA i.e. the SE is 1.7/sqrt(100)=0.17



How is this useful?

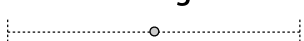
- We know the TA is 16.5 because that is the middle point of the high-schoolers' ages.
    - In practice, however, we do NOT know what the TA is!
- The given sample has a SA of 16.82 and SE of 0.17
- We can mathematically prove that with 95% probability, the true average (16.5) falls within two standard errors (2*0.17) around sample average (16.82),
    - I.e., 95% confidence interval (CI) [SA-2*SE, SA+2*SE]
- We know that the 95% CI, [16.82-2*0.17, 16.82-2*0.17], contains the TA with 95% probability!



We can easily construct 95% confidence intervals for many other measures of interest

EXAMPLE OF OTHER MEASURE

Sample Average

-2*SE  **+2*SE** where $SE = \frac{SD}{\sqrt{N}}$

Difference between two sample averages

-2*SE  **+2*SE** where $SE_{\text{diff}} = \sqrt{(SE_1)^2 + (SE_2)^2}$

Regression coefficient

-2*SE  **+2*SE** where $SE = \ldots$

Relationship between "95% confidence interval" and p-value

INTUITION

- The true average is unknown, but is inside the "95% confidence interval" with 95% probability

\$3.76

-2*SE **+2*SE**

"95% Confidence Interval"

- H0: 3.76 is no far away from 0 (i.e., expenditures of urban and rural customers show NO difference)

p-values AND CONFIDENCE INTERVALS

- Accpt H0: If the t-stat has a p-value ≥ 0.05 <==> "0" falls **inside** the 95% confidence interval
- Rej. H0: If the t-stat has a p-value < 0.05 <==> "0" falls **outside** the 95% confidence interval

How exactly we describe data and make statistical inferences from it depends on the types of variables we are trying to associate

STATISTICAL METHODS BY TYPE OF VARIABLES

Association between two numeric variables (e.g. age and income)

- Scatter plot
- Correlation (t-test to assess significance)
- Regression (t-test and F-test to assess significance)

Association between two non-numeric variables (e.g. gender and auto style)

- Cross-tabs (chi-square-test to assess significance)

Association between one numeric and one non-numeric variable (e.g. gender and income)

- Compare means of each group (t-test to assess significance)
- Regression with dummy (0-1) variable (t-test to assess significance)

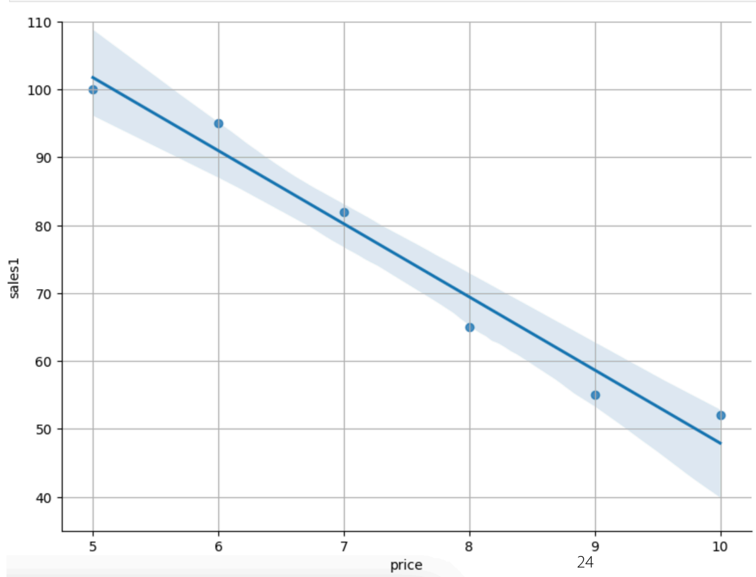
Inference in Practice: Association btw Two Numerical Variables

Case 1: Measuring the association between two numeric variables

METHOD 1: SCATTERPLOT

- Variable 1: "price"
- Variable 2: "sales1"
- In Python:

```
sns.lmplot(x="price", y="sales1", data=price_sales, ci=95, height=6, aspect=8/6)
plt.ylim(35, 110)
plt.grid(True)
plt.show()
```



Case 1: Measuring the association between two numeric variables

METHOD 2: CORRELATION COEFFICIENT

- Bounded between –1 and 1
- Indicates direction (negative or positive) and consistency of association
- Can assess statistical significance (using a t-test)
- Variable 1: "price"
- Variable 2: "sales1"
- In Python:

Correlation coefficient

```python
# correlation between price and sales1 with p value
from scipy.stats import pearsonr
corr, p_val = pearsonr(price_sales["price"], price_sales["sales1"])
print(f"Correlation between price and sales1: {corr:.3f}")
print(f"P-value: {p_val:.3f}")
```

```
Correlation between price and sales1: -0.983
P-value: 0.000
```

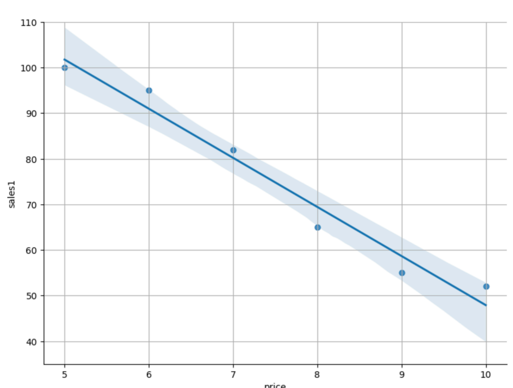
p-value for Hypothesis 0: " Price and Sales are uncorrelated, i.e. the correlation coefficient is 0"

Correlation coefficients can be misleading because they contain little information

EXAMPLE: SECOND SALES REGION

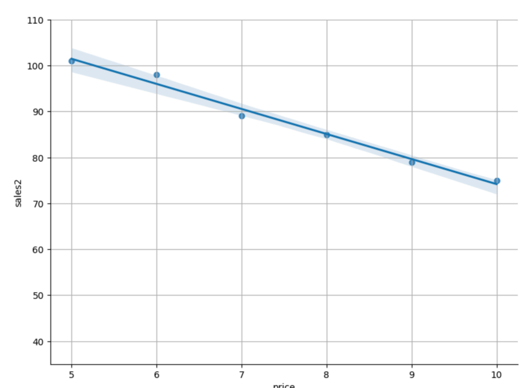
First sales region: "sales1"

```
Correlation between price and sales1: -0.983
P-value: 0.000
```



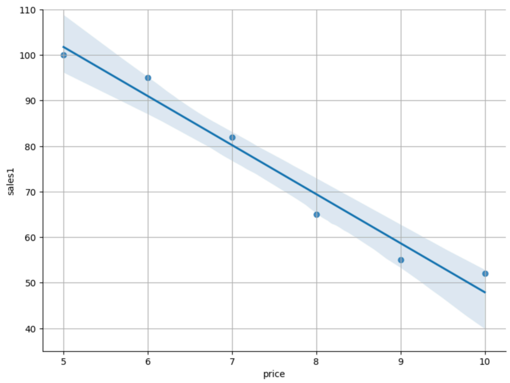
Second sales region: "sales2"

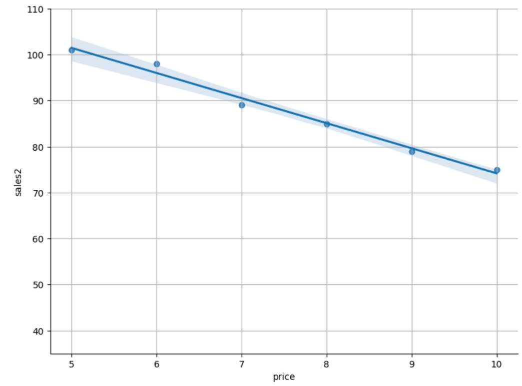
```
Correlation between price and sales2: -0.993
P-value: 0.000
```



Case 1: Measuring the association between two numeric variables

A BETTER SOLUTION: REGRESSION





- Regression describes the relationship between sales and price by "fitting the line" that best describes the data
- "sales = a + b• price": what is b and is it statistically significantly different from 0?
- Mathematically: Find the line that minimizes the sum of the squares of the vertical distances from the line to each data point
- Regressions capture more information about the data than correlation coefficient
- They are easier to interpret

Case 1: Measuring the association between two numeric variables

REGRESSION

```
# Perform linear regression
model1 = smf.ols('sales1 ~ price', data=price_sales).fit()
model2 = smf.ols('sales2 ~ price', data=price_sales).fit()
# Print the summary of the regression results
print(model1.summary())
print(model2.summary())
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    155.6190      7.824     19.891      0.000     133.897     177.341
price        -10.7714      1.017    -10.590      0.000     -13.595      -7.947
==============================================================================
Omnibus:                      nan   Durbin-Watson:                   1.906
Prob(Omnibus):                nan   Jarque-Bera (JB):                0.726
Skew:                      -0.017   Prob(JB):                        0.696
Kurtosis:                   1.297   Cond. No.                         35.2
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    128.7619      2.547     50.550      0.000     121.690     135.834
price         -5.4571      0.331    -16.479      0.000      -6.377      -4.538
==============================================================================
Omnibus:                      nan   Durbin-Watson:                   3.013
Prob(Omnibus):                nan   Jarque-Bera (JB):                0.368
Skew:                       0.474   Prob(JB):                        0.832
Kurtosis:                   2.241   Cond. No.                         35.2
==============================================================================
```

If one variables has a statistically significant effect on another variable, this does not mean that it is important

SIGNIFICANCE VS. IMPORTANCE: EXAMPLE OF SALES 3

price	sales3
10	97.0
9	97.4
8	98.0
7	98.1
6	99.0
5	100.0

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    102.5143      0.534    191.876      0.000     101.031     103.998
price         -0.5686      0.069     -8.186      0.001      -0.761      -0.376
==============================================================================
Omnibus:                      nan   Durbin-Watson:                   1.614
Prob(Omnibus):                nan   Jarque-Bera (JB):                0.335
Skew:                      -0.530   Prob(JB):                        0.846
Kurtosis:                   2.536   Cond. No.                         35.2
==============================================================================
```


Inference in Practice: Association btw Two Non-numerical Variables

Case 2: Associations between two non-numeric variables

METHOD: CROSS-TABS

- Situation: The Senate vote on Neil Gorsuch (Supreme Court Justice)
- Two Variables: Political Party and Vote

Rep.	52
Dem.	48

Yes	55
No	45

- Is there an association between a senator's political party and whether they voted for Neil Gorsuch?
- Cross-tabs allows you to determine the % who were Democrats and voted for Neil Gorsuch.

If there is NO relationship between party affiliation and vote (H0 is true), how many Dem./Rep. would vote Yes/No?

PREDICTED VOTES (assuming party and votes are unrelated): How would "NO/YES" be distributed across Dem. and Rep.?

	No	Yes	
Dem.			**48/100 => 48%**
Rep.			**52/100 => 52%**
	45	55	100

If there is NO relationship between party affiliation and vote (H0 is true), how many Dem./Rep. would vote Yes/No?

PREDICTED VOTES (assuming party and votes are unrelated): How would "NO/YES" be distributed across Dem. and Rep.?

	No	Yes	
Dem.	45*48%=21.6 48% are Dem., so 48% "No" should be Dem.	55*48%=26.4 48% are Dem., so 48% "Yes" should be Dem.	48 => 48%
Rep.	45*52%=23.4 52% are Rep., so 52% "No" should be Rep.	55*52%=28.6 52% are Rep., so 52% "Yes" should be Rep.	52 => 52%
	45	55	100

The actual vote seems to follow closer along party lines than the predicted vote

ACTUAL VOTES vs. EXPECTED VOTES (If H0 is true)

```
# Gorsuch chi-squared test
from scipy.stats import chi2_contingency

# Read the data
data_url = "https://tinyurl.com/gorsuch-txt"
gorsuch = pd.read_csv(data_url, delimiter='\t')

# Create a contingency table
contingency_table = pd.crosstab(gorsuch['party'], gorsuch['vote'])

# Perform the Chi-squared test
chi2_stat, p_value, dof, expected = chi2_contingency(contingency_table, correction=False)
```

```
print("Expected frequencies:", expected)
print("Observed frequencies:", contingency_table)
```

```
Expected frequencies: [[21.6 26.4]
 [23.4 28.6]]
Observed frequencies: vote   no  yes
party
Dem    45    3
Rep     0   52
```

Is the difference between the actual and the predicted vote (assuming party and vote are uncorrelated) significant?

We can use a chi-square test to determine whether the association between categorical variables is significant

TESTING FOR SIGNIFICANCE IN CROSS-TABS

– Compare the actual numbers with the expected numbers:

$$\frac{(45-21.6)^2}{21.6}+\frac{(0-23.4)^2}{23.4}+\frac{(3-26.4)^2}{26.4}+\frac{(52-28.6)^2}{28.6}=88.64$$

– The bigger the χ^2 test-statistic, the more the actual differs from the expected – meaning there is an association, that is, knowing one variable tells you something about the other variable

```
# Print the result in an easy-to-read format
print("Chi-square statistic:", chi2_stat)
print("p-value:", p_value)
print("Degrees of freedom:", dof)
```

```
Chi-square statistic: 88.63636363636361
p-value: 4.744862384520456e-21
Degrees of freedom: 1
```


Inference in Practice: Association btw One Non-numerical and One Numerical Variables

Case 3: Associations between numeric and non-numeric variable

METHOD 1: COMPARE MEANS OF EACH GROUP

- Non-metric variable splits the metric variable into groups
- Example: Executive pay and gender
  - Do female executive earn less than male executives?
  - Variable 1: salary
  - Variable 2: female (0-1)
- Calculate mean for each group
- Compare means using a t-test

```
# load salary data
data_url = "https://tinyurl.com/salary-experience-csv"
salary = pd.read_csv(data_url)

# Group by 'female' column and calculate the average salary
avg_salary = \
    salary.groupby('female')['salary'].mean().reset_index(name='avg_salary')
print(tabulate(avg_salary, headers='keys', tablefmt='grid',
               floatfmt=".2f", showindex=False))
```

```
+----------+--------------+
|   female |   avg_salary |
+==========+==============+
|     0.00 |    224224.61 |
+----------+--------------+
|     1.00 |    187585.79 |
+----------+--------------+
```


We use a "t-test" to determine whether the means are different between men and women

TESTING FOR DIFFERENCE BETWEEN MEANS

```
# Perform t-test of average salary between female and male
from scipy.stats import ttest_ind
female_salary = salary[salary['female'] == 1]['salary']
male_salary = salary[salary['female'] == 0]['salary']

t_stat, p_val = ttest_ind(female_salary, male_salary)

print(f"t-statistic: {t_stat:.3f}, p-value: {p_value:3f}")
```

```
t-statistic: -9.017, p-value: 0.000000
```

Case 3: Associations between metric and non-metric variable

METHOD 2: REGRESSION WITH DUMMY (0-1) VARIABLE

- Example: Executive pay and gender
  - Do female executive earn less than male executives?
  - Variable 1: salary
  - Variable 2: female (0-1)

```
# We can also regression model to test
# the significance of the difference

salary_reg = smf.ols('salary ~ female', data=salary).fit()
print(salary_reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.075
Model:                            OLS   Adj. R-squared:                  0.074
Method:                 Least Squares   F-statistic:                     81.30
Date:                Tue, 31 Dec 2024   Prob (F-statistic):           9.68e-19
Time:                        17:39:27   Log-Likelihood:                -12484.
No. Observations:                1000   AIC:                         2.497e+04
Df Residuals:                     998   BIC:                         2.498e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2.242e+05   2731.822     82.079      0.000    2.19e+05     2.3e+05
female     -3.664e+04   4063.340     -9.017      0.000   -4.46e+04   -2.87e+04
==============================================================================
Omnibus:                       14.827   Durbin-Watson:                   1.225
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                9.031
Skew:                          -0.025   Prob(JB):                       0.0109
Kurtosis:                       2.537   Cond. No.                         2.52
==============================================================================
```


Statistical Significance does not imply Causation; Also, what does "control" a variable mean?

Evidence suggests that heavier children have higher reading achievement

Correlation coefficient: 0.809
P-value: 0.000

Does higher weight cause children to read better?

A modified scatter plot highlights the problem

SCATTERPLOT WITH COLOR CODING