STAT 200

May 16 Ch. 1, 2

email: stat200@ugrad.stat.ubc.ca
WeBWork
Labs due on the day of lab
Lecture notes
2 written assignments

Randomness and its consequence, uncertainty
- pollsters —— results, distant from each other
- math & model → irl phenomena: statistics
- sample group: demographic matters, size matters
- e.g. smoking ⟷ health
  - prospective observational study. 1950
  - 40.000+ doctors
  - smoking → disease
  - clinical trials

e.g. COVID19 vaccines
- process uncertainties? → side effects etc.

- Monty Hall Problem

Data
- e.g. ages
  - numerical values, can be quantitative

e.g. handedness
- categorical variable

e.g. heights
- quantitative variable (continuous scale)

e.g. # of statistics course

e.g. phone area code: categorical → discrete

e.g. Rating of a movie (dependency of the continuous scale)

Statistics
- design of studies
- data collection
- summarizing and analyzing the data ✱
- interpreting the results
- drawing conclusions → made about specific phenomena (limited information)

→ May 16

Process of Investigation
- Collect data
- Examine data
- Interpret the results & draw conclusions
- e.g. Earthquakes
  - can categorize by year
  - can set up range, average → SD

Variable: characteristic of interest (Row: information corresponding to an individual, or an object or an experiment unit)
- Qualitative / Categorical
- Quantitative → can be ordered: ordinal variables
  - → measured on a numerical scale
    attach units

Understanding the Data
- Who · Where · Why
- What · When · How

Displaying data
- Categorical Variables (group similar things together)
  - (Relative) Frequency Tables
    → expressed in percentages / proportions
    add to 100%
  - Bar charts
    → Height ∝ Proportions
    → Same width
    } Area ∝ Height ∝ Proportions
    Legends, Labelled axes, Notes
  - Pie charts
    → can only introduce some # of categories
  - Contingency Tables
    → 2 categorical variable relationship
    → Breakdown of data

	Region (numbers are FREQUENCIES)											
Felt?	AB	BC	MB	NB	NL	NS	NT	NU	ON	PE	QC	YT
No	21	1682	2	40	3	5	119	223	77	1	362	190
Yes	0	15	0	30	1	0	1	0	12	0	28	2
Total	21	1697	2	70	4	5	120	223	89	1	390	192

} near communities?

→ May 16

A study looks at the relationship between diet type (high versus low cholesterol diet) and presence of coronary heart disease. Here are data collected on 23 individuals:

	Having heart disease?		Total
	Yes	No	
High cholesterol diet	11	4	15
Low cholesterol diet	2	6	8
Total	13	10	23

distribution of one variable collapsing the other variable

marginal distribution (frequency taken along the margins of the table)

conditional distribution
→ fix one variable, distribution for the other variable

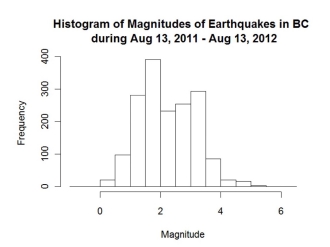
Variable association ?
- if not associated → similar percentages

Simpson's Paradox

May 17 Ch. 3.4

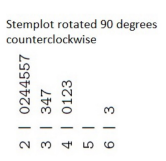
Displaying data

- Quantitative
  - Histogram

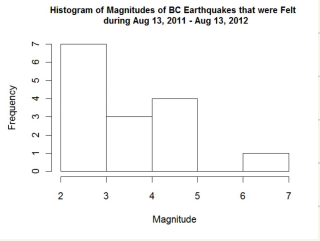


- changing width
  → different appearance
- no gaps between intervals

- Stem-and-leaf display
  - observation = stem + leaf
  - e.g. 2.4
    trailing digit : 4 (leaf)
    stem : 2
    137 cm
    trailing digit : 7 (leaf)
    stem : 13

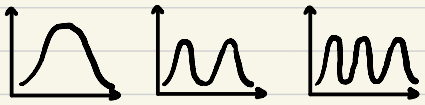


- Boxplot



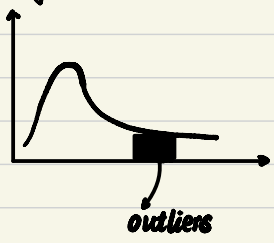
Describing distribution

1. Shape → unimodal, bimodal, multimodal

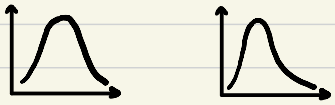


2. Center
3. Spread

Any outliers?



Symmetric vs. Skewed (skew tells tail)



Simpson's Paradox

- All data: one trend
- After proper grouping: trends reverse

	Large hospital	Small hospital
Total # surgeries	1000	300
# delayed discharge	130	30
% delayed discharge	130/1000=13%	30/300=10%

Large hospital

	Major surgery	Minor surgery
Total # surgeries	800	200
# delayed discharge	120	10
% delayed discharge	120/800=15%	10/200=5%

Small hospital

	Major surgery	Minor surgery
Total # surgeries	50	250
# delayed discharge	10	20
% delayed discharge	10/50=20%	20/250=8%

→ May 17

Measures of Center

1. Mean (arithmetic average)

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$, y_i is ith observation

2. Median
   - arrange data in ascending order
   - odd: median = $\frac{n+1}{2}$ th observation
     even: median = average of $\frac{n}{2}$ and $\frac{n+1}{2}$ th observations

Measures of Spread

1. Range
   max − min
2. IQR (interquartile range)
   range encloses the middle 50% observations
   "pth percentile" $(0<p<100)$: p% fall below it, $(100-p)$% fall above it
   Quartiles (Q_1: 25th, Q_2: 50th, Q_3: 75th)
   $IQR = Q_3 - Q_1$
   If odd, include median in both Q_1 and Q_3

y_1, y_2, y_3, y_4, y_5
- y_3: median
- $y_1, y_2, y_3 \rightarrow Q_1$: y_2
- $y_3, y_4, y_5 \rightarrow Q_3$: y_4

y_6
- $\frac{y_3+y_4}{2}$: median
- $y_1, y_2, y_3 \rightarrow Q_1$
- $y_4, y_5, y_6 \rightarrow Q_3$

3. Variance and Standard deviation

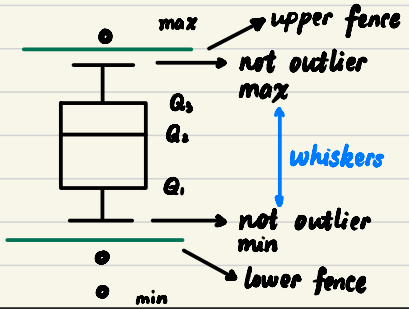
$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$ → sample (unbiased) ↑ corrects the bias

$$\sigma^2 = \frac{\cdots}{n}$$ → theoretical (biased)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$ (on a standard scale)

- non-negative
- more spread, larger variance
- SD same unit as data
- $s^2 = s = 0$ ← all observations are the same

5-number summary
- min, Q_1, Q_2, Q_3, max → box-plot
- Outlier: $> Q_3 + 1.5 \times IQR$
  $Q_1 - 1.5 \times IQR$



distribution: histogram

→ May 17

Sensitivity to Outliers

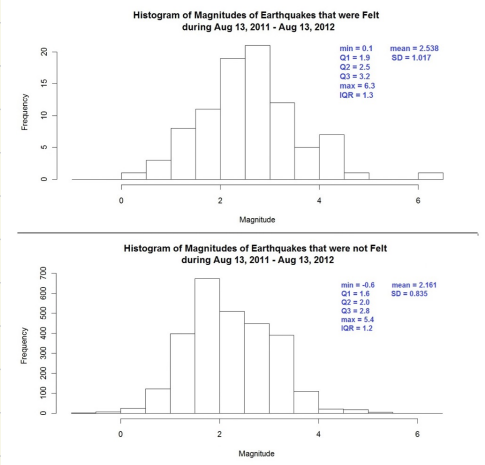
Sensitive	Not sensitive
• mean	• median
• range, variance, SD	• IQR

What to use?

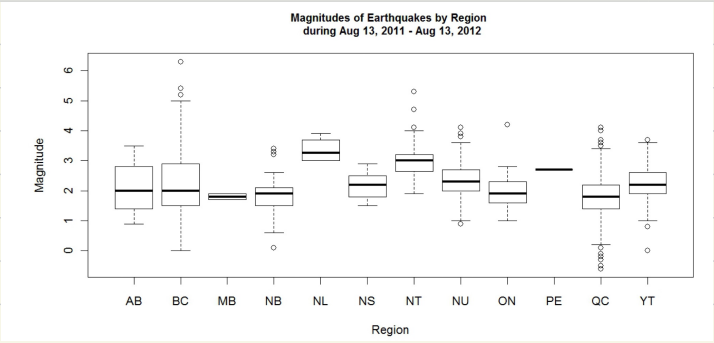
- range: crude measure
- symmetric: mean, variance, SD
- skewed: median, IQR

Comparing Distributions with Histogram

- always same scale, same bins



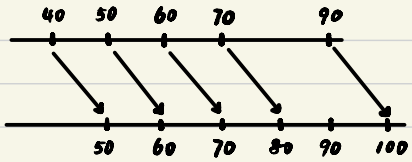
Comparing ... Boxplots



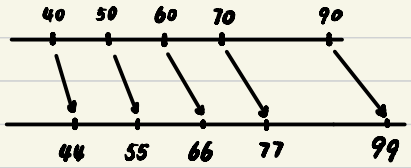
May 19 Ch. 5, 6, 7, 8

Shifting, Scaling ⟶ mean, SD & summary ··· ?

1. Shifting
   - add constant c to each observation
   - measure of center : $+c$
   - measure of spread : unchanged

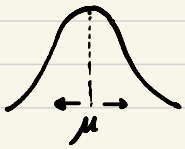


2. Scaling
   - multiply constant c
   - measure of center & spread : $\times c$



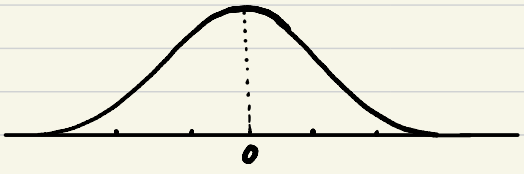
Standardizing
- 2 different scale comparison
- z-score : how far is observation from center

$$z = \frac{y - \bar{y}}{s}, \quad \bar{y}\ (\text{mean}),\ s\ (\text{SD})$$

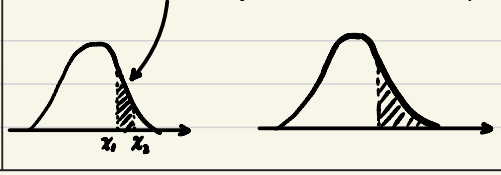


$N(\mu, \sigma)$

$z \begin{cases} k : > \mu \\ 0 : = \mu \\ -k : < \mu \end{cases}$



Normal model (quantitative variables)
- what percentage is [range of variables] in [observations]?



- bell shaped, unimodal
- perfectly symmetric about μ
- spread : SD is σ
- denoted by $N(\mu, \sigma)$

} parameters ← from model
↕ from data
$\bar{y}, s$: statistics

→ May 19

Standardizing values from the Normal Model

$$z = \frac{\text{observation - mean}}{\text{standard deviation}} = \frac{y-\mu}{\sigma}$$

68 - 95 - 99.7 rule

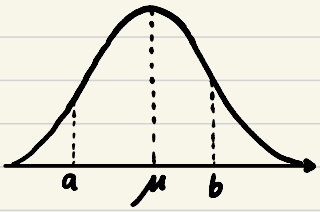
$|y-\mu| \le \sigma$: 68%

$|y-\mu| \le 2\sigma$: 95%

$|y-\mu| \le 3\sigma$: 99.7%

Finding areas under the Normal Model

1. Sketch



2. Obtain value of area from table OR Software it

1. Scores on a standard IQ test for the 20 to 34 age group follow approximately the Normal model with mean $\mu = 110$ and standard deviation $\sigma = 25$.

(a) What percentage of people aged 20 to 34 have IQ scores below 160?

$z = \frac{160-110}{25} = 2 \rightarrow \le 2\sigma \rightarrow 97.5\% \quad (95\% + \frac{5\%}{2})$

(b) What percentage have scores between 90 and 120?

$z_1 = \frac{90-110}{25} = -0.8$

$z_2 = \frac{120-110}{25} = 0.4$

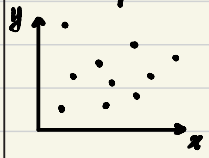
$\int_{-\infty}^{0.4} f(x)dx - \int_{-\infty}^{-0.8} f(x)dx$

$\rightarrow -0.8\sigma \le y \le 0.4\sigma \rightarrow$ table $(0.6554 - 0.2119 = 0.4435)$

(c) How high is the IQ such that only 0.15% of the group fall above?

$\int_{-\infty}^{t} f(x)dx = 0.9985 \rightarrow$ table $\rightarrow t \approx 3$

Scatterplots, Association, & Correlation



$(-1 \le r \le 1)$

Linearity!

→ May 19

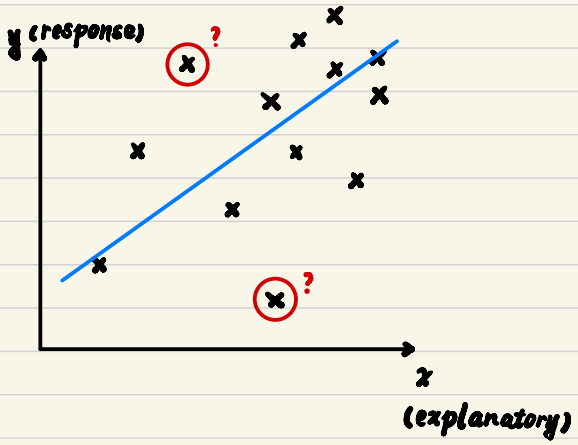
Human intelligence ⟷ IQ (Quantitative Variable)
Class skipping ⟷ Grades

Scatter plot
- 2 quantitative variables
- x_i, y_i
- $(\bar{x}, \bar{y})$: mean-mean point
  - center of data cloud
  - line of regression always cross $(\bar{x}, \bar{y})$
- +, - directions
- Linear, Non-linear
- Strong / Weak / No relationship
- Outlier ?
- explanatory variable : x-axis
  response variable : y-axis
) some influence

May 23

Types of Patterns of a Scatterplot

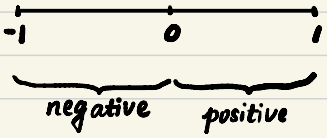
1. Direction
   - Positive (x, y same direction)
   - Negative (… opposite …)
2. Form
   - Linear
   - Non-linear
3. How scattered?
   - Strong / weak / no relationship
4. Any outliers?



correlation ≠ causation

Correlation

- degree of linear association between 2 **quantitative variables**
- correlation correlation (r): measure strength of **ONLY** linear correlation



- sensitive to outliers
- { swapping x & y
  + c or × c (positive) } → does not change r

Calculating r

1. standardize x & y

$$z_x = \frac{x-\bar{x}}{s_x}, \quad z_y = \frac{y-\bar{y}}{s_y}$$

2. $$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum \left(\frac{x-\bar{x}}{s_x} \cdot \frac{y-\bar{y}}{s_y}\right)}{n-1}$$

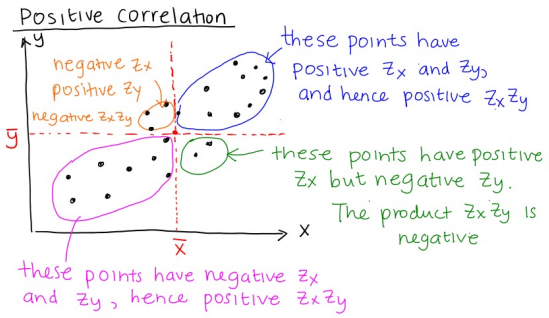
Covariance (not standardized)

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$r = \frac{cov(X,Y)}{s_x s_y}$$

- variance: same variable $(x_i-\bar{x})^2$

Concept on correlation coefficient



The correlation coefficient is obtained by adding these products $z_x z_y$ (then by dividing n-1). Note that here there are more positive $z_x z_y$ than the negative ones. Hence the sum of all $z_x z_y$'s will be positive. The value of r is positive.

Exercise: think about why r is negative for a negative correlation!

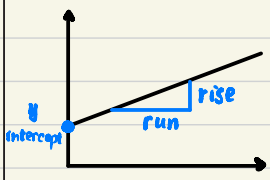
→ May 23

Linear Regression

- Describe linear relationship between 2 quantitative variable
- Predict on the response y given explanatory x
- Model

y = intercept + (slope) × (x)

$y = b_0 + b_1 x$, b_0 = intercept, b_1 = slope



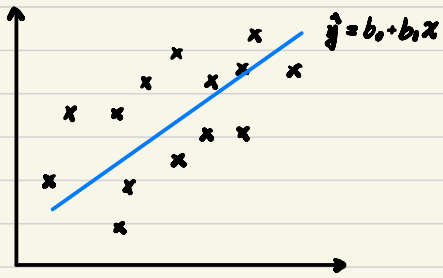
$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x}$$

Regression Line

- $b_1 = \frac{r s_y}{s_x}$ · r (correlation coefficient)
- $b_0 = \bar{y} - b_1 \bar{x}$
- line always passes through $(\bar{x}, \bar{y})$
- model relating x & y is:

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

→ predicted value



Beware of extrapolation → (outside the observed range)

- Grades vs. Class skipped
- Skipping 30 classes? → unreasonable

b_1:

- an increase of 1 S_x in x is associated with a change of $r \times S_y$ in $\hat{y}$
- a unit increase in x is associated with b_1 that much change in $\hat{y}$

b_0:

- predicted value when $x = 0$

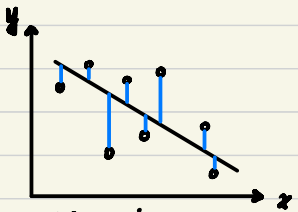
Residuals (e)

- $e = y - \hat{y}$

$\sum (y_i - \hat{y}_i) = 0$

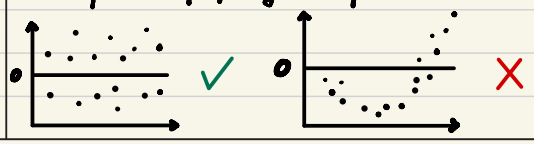
- minimize $\sum (y_i - \hat{y}_i)^2$

→ ordinary least squares regression



- residual plots

→ if done properly, no pattern.



Association vs. Causality

$\not\rightarrow$

does not imply!

- lurking variable: a 3rd variable that associates with both x & y

→ May 23

Fitting linear model

- x & y has sufficient linearity
  (observation transformation : log, √)
- Reality check : do the values make sense
- Attention to outliers → on the end affects more (exclusion of data ?)
- Do not extrapolate

May 24

Sample Surveys

- complete collection of individuals under study : population
  - costly, inefficient
- sample ⊆ population
  - non-representative: biased
- parameter : numerical characteristic of population
  - counterpart of a sample ⟶ statistic
  - part of a model for a population : population parameter
  - statistics —estimate→ population parameters
    (m, s) ⟷ (μ, σ)

Randomization

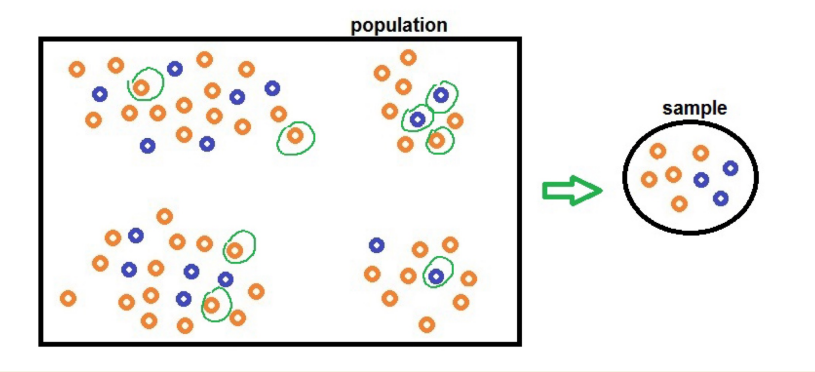
- Randomness : give samples that have population characteristics
  - Prevent non-representativeness
- Sample size matters : actual number of individuals/subjects is important { not population size; not fraction
  - Large enough sample ⟶ representative

Sampling methods

- Sampling frame: every subject that CAN (→ SHOULD) be sampled, define population clearly
- Sampling variability: difference in characteristics from sample to sample
- Sample size ↑, variability ↓

identify who's missing

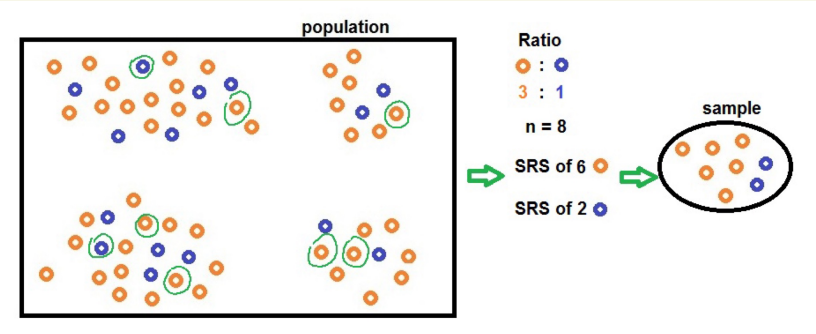
1. Simple random sampling (SRS)
   - each subject : equal chance of being selected
   - each possible sample of size n is equally likely



→ May 24

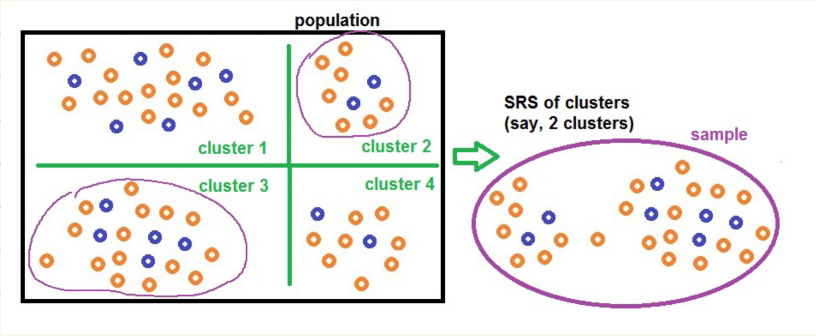
2. Stratified sampling

- population ⟶ strata (subset with same particular characteristic)
- SRS drawn within each stratum ⟶ stratified sample



3. Cluster sampling

- population naturally divided into groups or clusters
- SRSs from all available clusters are obtained (if all individuals from selected clusters are included in sample ⟶ final sample: one-stage cluster sample)



4. Multistage sampling

- 2-stage cluster sampling

5. Systematic sampling

- selecting every k^{th} individual from the sampling frame
- works if list has no hidden order

Bad sampling

- Undercoverage
  - completely excludes / underrepresents certain kinds of individuals
- Convenience sampling
  - based on easy availability & accessibility
- Voluntary response bias
  - individuals with strong opinions tend to respond more often ⟶ overrepresented

May 26

Bad sampling

- non response bias ⟷ voluntary response bias
- response bias
  - response influenced by how questions are phrased / worded
  - misunderstanding of a question
  - unwillingness to disclose the truth

Probability and Random variables

- Sample space: S
  - set of all possible outcomes of a random phenomenon

S

head	tail

- Event: outcome(s) from a random phenomenon
  - rolling a 3 on a die
  - denote by uppercase letter
- P(A): probability that an event A will occur
  - $0 \leq P(A) \leq 1$ P(A = head) = 0.5
    - P(A) = 0 : impossible
    - P(A) = 1 : certain
    - ↗ P(A), ↗ likely A occurs
  - ∑ Probability of all non-overlapping events in S = 1
- True positive: Spam, classified as spam
  True negative: Ham, classified as ham
  False positive: Ham, classified as spam
  False negative: Spam, classified as ham
  - Positive: classified as spam
  - Negative: classified as ham
  - True: correct classification
  - False: incorrect classification
  - False positive rate =

$$\frac{\#\text{ham classified as spam}}{\#\text{ham}}$$

False negative rate =

$$\frac{\#\text{spam classified as ham}}{\#\text{spam}}$$

Indepence of events

- A & B are indepent if:
  P(B given A) = P(B) ⟶ P(B|A)
  P(A given B) = P(A) ⟶ P(A|B)
- if A&B are independent, the probability that A&B occur together is:
  P(A&B) = P(A) × P(B)

May 30

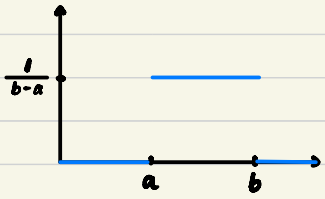
Random Variables

- discrete random variable: a numeric quantity of an outcome from a random phenomenon
  - → the variable can take on come from a countable set of values
- set of possible values + associated theoretical probability = probability model
- continuous RV: uncountable set of values ⟶ denoted by density curves

- normal random variables

- Uniform RV ($X \sim U(a,b)$)

runif(…) : random uniform



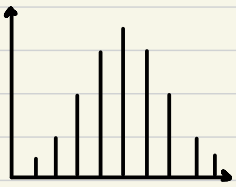
mean: $\frac{a+b}{2}$

sd: $\sqrt{\frac{(b-a)^2}{12}}$

Mean and variance of RVs

- mean: long run average of the observed values ⟶ weighted sum
- variance: how spread out the observations are

Binomial Model



- Binomial experiment
  - (n) identical trials
  - dichotomized into 2 complementary categories "success" & "failure"
  - $P(X=\text{success}) = p$, $P(X=\text{failure}) = 1-p = q$
  - each trial is independent
- Denoted as $X \sim Bin(n,p)$
- for $X \sim Bin(n,p)$:

$$P(X=x) = C_n^x p^x q^{n-x}, \quad C_n^x = \frac{n!}{x!(n-x)!}$$

($C_n^x = \binom{n}{x}$)

- mean: np
- variance: npq
- sd: $\sqrt{npq}$

- overestimate > underestimate

May 31

Sampling Distribution Models

- population ⟷ sample
  parameter ⟷ statistic

Sampling Distribution of Proportions (Percentage)

- True proportion (population, parameter): p $(0<p<1)$
- Sample proportion:

$$\hat{p} = \frac{\#\text{ individuals sampled, with characteristic}}{\text{sample size } n} \longrightarrow \text{estimate } p$$

- Reliability of $\hat{p}$ estimating p?
  - center & spread of sample proportion values?

Sampling distribution of $\hat{p}$

- $\mu(\hat{p}) = p$
- $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}}$ or $\sqrt{\frac{p(1-p)}{n}}$

$\Big\}\ \hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ ⟷ $n=10$ (in class)

- samples are randomly drawn
  sample size ≤ 10%
  individual values are independent
  sample size be large ($np \geq 10$ & $nq \geq 10$)

Sampling Distribution of Means

- Distribution of sampled mean
- center: true population mean ⟶ ↑sample size, closer to true mean
- μ: parameter
  $\bar{y}$: statistic } estimate
- $\begin{cases} \mu(\bar{y}) = \mu \\ \sigma(\bar{y}) = \frac{\sigma}{\sqrt{n}} \end{cases}$ (variability of sample mean around true mean)

$$\therefore \bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

σ ↓

S

Central Limit Theorem (CLT)

- $y_1, y_2, y_3, \cdots, y_n$ be independent (random sample) $\xrightarrow{\text{sufficiently large}}$ $\bar{y}$: (approx) Normal model
  - randomness, independence, large
- if sample comes from normal, sample mean is always normal

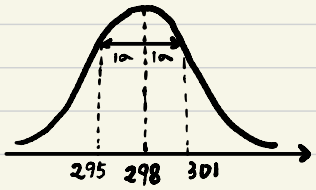
June 2

Ch 14

Exercise #3

A bottling company uses a filling machine to fill plastic bottles with cola. A bottle should contain 300 mL. In fact, the contents vary according to the Normal model with mean 298 mL and standard deviation 3 mL.

(a) What is the probability that an individual bottle contains less than 295 mL?



By empirical rule,

$z_{295} = -1\sigma$

68% will be between 295mL ~ 301mL

$\therefore$ By symmetry, $P(X < 295) = \frac{1 - 0.68}{2} = 0.16$

(b) What is the probability that the mean content of bottles in a six-pack is less than 295 mL?

$n = 6$,

mean of $\bar{y} = 298\text{mL}$

SD of $\bar{y} = \frac{3\text{mL}}{\sqrt{6}} \approx 1.225\text{mL}$

$\therefore z = \frac{295 - 298}{1.225} \approx -2.449$

From table: $P(X < 295) \approx 0.0071$

Binomial vs. Normal

- Sample size > 70, almost indistinguishable

Dummy Writen Hw

Q1

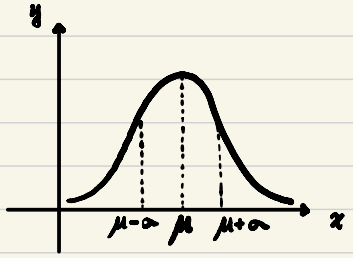
Student name: Wenyou (Tobias) Tian

Student number: 44045599

Q2

Major: Not declared, planning to pursue Computer Science and Economics

Q3



June 6

Confidence Intervals for Proportions

- p: population proportion
  $\hat{p}$: sample proportion

$\longrightarrow \hat{p} \sim N(\mu(\hat{p}) = p,\ \sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}})$

- no repeated sample, estimate p with $\hat{p}$
  variability of $\hat{p}$:
  $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

By empiral rule, and using $SE(\hat{p})$:

68%: $\hat{p}$ in $p \pm SE(\hat{p})$	$\hat{p} \pm SE(\hat{p})$	68% captures p
95%: $\hat{p}$ in $p \pm 2SE(\hat{p})$ $\longrightarrow$	$\hat{p} \pm 2SE(\hat{p})$	95%
99.7%: $\hat{p}$ in $p \pm 3SE(\hat{p})$	$\hat{p} \pm 3SE(\hat{p})$	99.7%

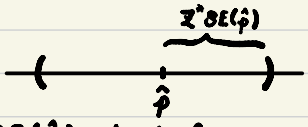
- Varying intervals ($SE(\hat{p})$): confidence intervals for population proportion p
  one-proportion z-intervals
- General formula for CI for p:

$\hat{p} \pm z^* SE(\hat{p}) \longleftrightarrow \hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}$

- Confidence level (C): probability that a CI encloses p
  common: 90%, 95%, 99%

upper tail area $= \frac{100\% - CI}{2} \longrightarrow z^* = z$-score

- Margin of error (ME) — precision / certainty
  $ME = z^* SE(\hat{p})$
- Assumptions:
  - Randomness, independence, sufficiency
- Properties of CI



Sample size ↑ ×2, Standard error ↓ ÷$\sqrt{2}$

- $SE(\hat{p})$ built for parameter p

NOT Probability

$\longrightarrow$ We are C% confident that p is between <interval>

Sample size determination

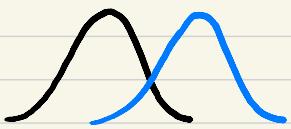
- appropriate sample size n $\longrightarrow$ certain precision for parameter estimation

$\because ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \therefore n = \frac{(z^*)^2 \hat{p}(1-\hat{p})}{ME^2}$

- Largest sample size: ① $\hat{p}(1-\hat{p}) = (0.5)(0.5)$ ② ME specified

maximum sample size

Testing Hypotheses (Significance Test)



- actual change OR sampling variation

- hypothesis: a statement/claim about a parameter
- null hypothesis (H_0): statement about the value of parameter whose general form is:

$$H_0: p = p_0$$

(p: parameter; p_0: some value)

$0 < p_0 < 1$

- alternative hypothesis (H_A): opposes H_0

$$\begin{cases} H_A: p \neq p_0 & \text{(2-tail test)} \quad \text{2-side} \\ H_A: p > p_0 & \text{(right tail test)} \quad \text{1 side} \\ H_A: p < p_0 & \text{(left tail test)} \quad \text{1 side} \end{cases}$$

- hypothesis test: one-proportion z-test

June 7

Exercise

The overall rate of defects in a manufacturing process has been 6%. After experimenting with a new manufacturing process, the quality control department is interested in testing if the new process significantly reduces the defective rate. A random sample of 300 machines produced using the new process revealed that 15 are defective. Do the data provide sufficient evidence that the defective rate is reduced?

A hypothesis test is to be carried out. What is the parameter of interest? What are the hypotheses?

parameter : proportion of defective widgets

null hypothesis — $H_0 : p = 0.06$, the true proportion is the same as the previous

alternative hypothesis — $H_A : p < 0.06$, the true proportion is less than the previous

- Starting with : null is true

  assuming this

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

$np_0 \geq 10$, $n(1-p_0) \geq 10$

Exercise (cont'd): Write down the null model for the hypothesis test.

Let $\hat{p}$ be the proportion of defects in the sample,

p_0 be the proportion in the null distribution

Null model:

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

where $p_0 = 0.06$ and $n = 300$

- Determine how far away $\hat{p}$ is from p_0 by z-score

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

→ test statistic

→ observed difference large (data is incompatible with null)

Exercise (cont'd): Compute the test statistic for the hypothesis test. Recall that a random sample of 300 machines produced using the new process revealed that 15 are defective. The defective rate prior to the implementation of the new process has been 6%.

$$Z_0 = \frac{\frac{15}{300} - 0.06}{0.014}$$

$$= -0.729$$

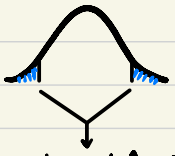
not extreme

→ June 7

- compute p-value to evaluate how unusual an observed difference is when H_0 is true

(1) $H_A: p \neq p_0$ (2) $H_A: p < p_0$ (3) $H_A: p > p_0$

p-value = shaded area



observed $\hat{p}$ (either one)

Exercise (cont'd): Compute the P-value of the hypothesis test.

$Z_0 = -0.729$, probability ? ⟶ 0.2327 is our p-value

0.01 for a 99% significance level ∵ $0.2327 > 0.01$

∴ fail to reject H_0

- small P-value suggests observation is unlikely due to sampling variation if H_0 is true
- P-value: conditional probability given H_0 is true, do not misinterpret P-value to be $P(H_0 \text{ is true})$
- Significance level (α) $\alpha = 0.01, 0.05, 0.1$
  - P-value $< \alpha$: reject H_0, test is significant at α level
    $\geq \alpha$: do not reject H_0
- only (do not) reject H_0

Exercise (cont'd): Do we reject the null hypothesis at the 10% significance level? What is the conclusion?

fail to reject

1. Identify the population and the parameter that we want to test a hypothesis on.
2. Set up the hypotheses: Null and Alternative.
3. Compute a test statistic based on the data.
4. Compute the P-value.
5. Draw conclusion based on the P-value.

- Type I, II error
  - I: rejecting H_0 when H_0 is true. probability is α
  - II: failing to reject H_0 when H_0 is false

		True State of Nature	
		H_0 is true	H_A is true
Decision	Reject H_0	Type I error	Correct decision
	Do not reject H_0	Correct decision	Type II error

→ June 7

Exercise (cont'd): Identify the Type I and Type II errors of the hypothesis test.

I: if we had rejected H_0, we could have done so incorrectly

II: we failed to reject H_0, we could be wrong - maybe the proportion of defects did decrease

June 13

Population Parameter	Sample Statistic	Standard Deviation of Statistic	Standard Error of Statistic
Population Proportion p	Sample Proportion $\hat{p}$	$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$	$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Population Mean μ	Sample Mean $\bar{y}$	$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$	$SE(\bar{y}) = \frac{s}{\sqrt{n}}$

Inference about Mean

- Confidence interval for parameter:
  statistic ± critical value (standard error)
- Confidence level C for μ is:

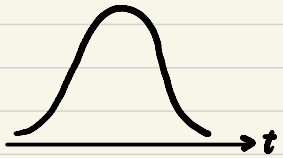
$$\bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

→ critical value, obtained from t-table

confidence interval for μ is called the <u>one-sample t-interval</u>

t-model

- one model parameter — degrees of freedom (df) ⟶ determines shape of model, given by $n-1$

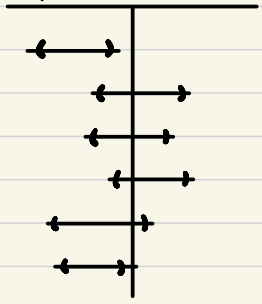


- eg. $n = 51$,
  $df = 51 - 1 = 50$,
  want 95% CI. $t = 2.009$

- Assumptions
  - random, independent, <10%, "near Normal"
- If σ is known:

$$\bar{y} \pm z^* \frac{\sigma}{\sqrt{n}}$$

confidence intervals



→ June 13

Exercise #1

A student in the STAT 200 class randomly selected 25 weekdays and timed how long the 8am #99 bus took to travel from Broadway Station to UBC. His data gave an average travel time of 37 minutes and standard deviation of 2.5 minutes.

Use the student's data to construct a 90% confidence interval for the true mean travel time between Broadway Station and UBC for the 8am #99 bus on weekdays. Also interpret the confidence interval in the context of this question.

$n = 25$, $df = n - 1 = 24$,
with 90% CI, $t^*_{n-1} = 1.318$

$\therefore \bar{y} \pm t^*_{n-1} \frac{s}{\sqrt{n}} \longleftrightarrow 37 \pm 1.318 \times \frac{2.5}{\sqrt{25}}$

$\downarrow$

37 ± 0.659

Hypothesis testing for μ

- Null — $H_0: \mu = \mu_0$ ⟶ one-sample t-test, $t_{n-1} = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$
- Alternative — $H_A: \mu \neq \mu_0$
  $H_A: \mu > \mu_0$
  $H_A: \mu < \mu_0$
  one-sample z-test, $z_0 = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
- p-value : sum of areas at tails

Exercise #1 (cont'd)

A student in the STAT 200 class randomly selected 25 weekdays and timed how long the 8am #99 bus took to travel from Broadway Station to UBC. His data gave an average travel time of 37 minutes and standard deviation of 2.5 minutes.

Do the student's data provide sufficient evidence that the true mean travel time is longer than 35 minutes? Carry out an appropriate hypothesis test at the 10% significance level.

Parameter of interest: average travel time

H_0: $\mu = 35$
H_A: $\mu > 35$

Is the population standard deviation known?
We shall use t-test *model.*

$t_0 = \frac{37 - 35}{\frac{2.5}{\sqrt{25}}} = 4$

p-value = 0.00026 3454

Reject H_0 ⟶ there is evidence that $\mu > \mu_0$

→ June 13

Exercise #2

To assess the accuracy of a laboratory scale, a standard weight which is known to weigh 1 gram is repeatedly weighed a total of 1000 times and the sample mean of the weighings is found to be 1.003 grams. Suppose the scale readings have an unknown mean μ and a known standard deviation σ=0.01 gram. Do the results from the 1000 weighings of the 1-gram standard weight suggest that the laboratory scale is not accurate? Carry out a hypothesis test at the 1% significance level.

parameter of interest: laboratory scale weight, $\alpha = 0.01$ → $z^* = 2.5758$

$H_0: \mu = 1.000$

$H_A: \mu \neq 1.003$

$\therefore z_0 = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1.003 - 1.000}{\frac{0.01}{\sqrt{1000}}} \approx 9.4868$

$\therefore 9.4868 > 2.5758$

Reject H_0

Comparing Means between 2 independent groups

- $y_{11}, y_{12}, y_{13}, \cdots, y_{1n_1}$ ← mean μ_1, sd σ_1

  $y_{21}, y_{22}, y_{23}, \cdots, y_{2n_1}$ ← mean μ_2, sd σ_2

Sampling distribution of mean differences

- $\bar{y}_1 - \bar{y}_2$ —estimate→ $\mu_1 - \mu_2$
- $\bar{y}_1 - \bar{y}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

→ $(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2)$, $df = \min(n_1 - 1, n_2 - 1)$

↓

two-sample t-interval

- Assumptions

→ June 13

Hypothesis tests

- $H_0: \mu_1 - \mu_2 = \Delta_0$ vs.
  $H_A: \mu_1 - \mu_2 \neq \Delta_0$
  $H_A: \mu_1 - \mu_2 > \Delta_0$
  $H_A: \mu_1 - \mu_2 < \Delta_0$

  for $\Delta_0 = 0$, $H_0: \mu_1 = \mu_2$ vs.
  $H_A: \mu_1 \neq \mu_2$
  $H_A: \mu_1 > \mu_2$
  $H_A: \mu_1 < \mu_2$

- test statistic

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

- Reject if p-value $< \alpha$

  Do not reject if p-value $\geq \alpha$

June 14

Exercise

The headmaster of an elementary school is interested in comparing two methods of teaching reading. He randomly selects two groups of students in his school and assigns each student to one of the two teaching methods for a 6-month period. At the end of the trial, each student writes a reading comprehension test, and the test scores are summarized in the following table.

	Method I	Method II
Number of students per group	41	42
Sample mean	75	67
Sample variance	52	71

Construct a 90% confidence interval for the difference in the true mean scores on the comprehension test between the two teaching methods.

$df = \min(41-1, 42-1) = 40$, 10% significance level

$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{52}{41} + \frac{71}{42}} \approx 1.72$

$8 \pm t^*_{40,0.1} \times 1.72 \longrightarrow 8 \pm 2.241$

(a) Do the data suggest that the true mean test score of students receiving Method I is higher than that of Method II by more than 5 points at the 10% significance level?

$\Delta_0 = 5 \longrightarrow H_0: \mu_1 - \mu_2 = 5$, $\alpha = 0.1$

$H_A: \mu_1 - \mu_2 > 5$

Test statistic:

$t_0 = \frac{75 - 67 - 5}{SE(\bar{y}_1 - \bar{y}_2)} \approx 1.744$ (p-value ≈ 0.044)

$t^*_{40,0.1} = 1.303$ (critical value)

$\longrightarrow 1.744 > 1.303 \longrightarrow$ reject $H_0 \longrightarrow$ sufficient evidence $\mu_1 - \mu_2 > 5$

(b) Repeat (a) if the headmaster wants to test whether the true mean test score is different for the two methods.

$\Delta_0 = 0 \longrightarrow H_0: \mu_1 - \mu_2 = 0$, $\alpha = 0.1$

$H_A: \mu_1 - \mu_2 \neq 0$

Test statistic:

$t_0 = \frac{75 - 67}{SE(\bar{y}_1 - \bar{y}_2)} \approx 4.651$

$t^*_{40,0.05} = 1.684$

$\because 4.651 > 1.684$

reject $H_0 \longrightarrow$ sufficient evidence $\mu_1 - \mu_2 \neq 0$

→ June 14

Comparing Means between 2 paired groups

- dependent / paired samples

  e.g. IQ of twins

  → look at within-pair difference

The general data structure for paired data will look like this:

Pair	Sample #1	Sample #2	Difference d
1	y_{11}	y_{12}	$d_1 = y_{12} - y_{11}$
2	y_{21}	y_{22}	$d_2 = y_{22} - y_{21}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	$d_n = y_{n2} - y_{n1}$
Sample mean	$\overline{y}_1$	$\overline{y}_2$	$\overline{d}$
Sample SD	s_1	s_2	s_d

Here, y_{ij} refers to the jth observation of the ith pair. There are a total of n pairs ($i = 1, 2, ..., n$) and each pair has two observations ($j = 1, 2$).

$\bar{d} \sim N(\mu_d, \frac{\sigma_d}{\sqrt{n}})$

↓

$\bar{d} \pm t^*_{n-1} \cdot \frac{s_d}{\sqrt{n}}$ (one sample t interval for population mean difference)

- Assumptions

  pairs are random
  independent
  ≤ 10% pop.
  σ unknown

- t-test

  $H_0 : \mu_d = \Delta_0$

  vs.

  $H_A : \mu_d \neq \Delta_0$

  $H_A : \mu_d > \Delta_0$

  $H_A : \mu_d < \Delta_0$

  test statistic:

  $$t_0 = \frac{\bar{d} - \Delta_0}{\frac{s_d}{\sqrt{n}}}$$

→ June 14

Exercise

To compare a new synthetic drug with an existing drug used to reduce eye pressure (in $mmHg$) in glaucoma (an eye disease), seven patients infected with the disease were treated with both drugs, one eye with the new drug and one with the existing drug. The reduction in pressure in each eye was then recorded and tabulated in the following.

Patient	Existing Drug	New Drug	
1	3.5	2.6	-0.9
2	2.6	2.8	0.2
3	3.0	3.1	0.1
4	1.9	2.4	0.5
5	2.9	2.9	0
6	2.4	2.2	-0.2
7	2.0	2.2	0.2

Assumptions ✓

$\bar{d} = \frac{-0.1}{7} \approx -0.01429$

$S_d \approx 0.445$

(a) Construct a 90% confidence interval for the mean difference in reduction in eye pressure between the two drugs. State any assumption(s) you have made.

$df = 7-1 = 6, \quad \alpha = 0.1$

$t^*_{6,0.1} = 1.943$

$\therefore \bar{d} \pm t^*_{6,0.1} \cdot \frac{S_d}{\sqrt{7}} \longleftrightarrow -0.01429 \pm 1.943 \times 0.167$

0.324481

(b) Do the data suggest a difference in the mean reduction in eye pressure between the two drugs? Use α=0.10.

$H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

Test statistics

$t_0 = \frac{-0.01429}{0.167} \approx -0.08557$

$\alpha = 0.10$

↓

$t = 1.943$

$|t_0| < t$

↓

fail to reject

↓

sufficient evidence : not different

June 16

Comparing counts

- Formulate hypothesis test for testing if there is an association between 2 categorical variables

	Having heart disease?		Total
	Yes	No	
High cholesterol diet	11	4	15
Low cholesterol diet	2	6	8
Total	13	10	23

- exactly same ⟶ independent / no association
- very different ⟶ dependent / associated
- between ⟶ ?

(conditional proportions)

Example

A survey involving 1772 adults gave the following data on alcohol consumption patterns by marital status.

	Drinks per month			
	Abstain	1–60	Over 60	Total
Single	67	213	74	354
Married	411	633	129	1173
Widowed	85	51	7	143
Divorced	27	60	15	102
Total	590	957	225	1772

Question of interest: are marital status and alcohol consumption independent?

Testing for independence (χ^2 testing)

- H_0 : not associated
  H_A : associated
- row ⟶ r categories
  column ⟶ c categories
  total # : n
- Assume H_0 is true:

expected counts

	Handedness		Total
	Right-handed	Left-handed	
Male	? 54	? 6	60
Female	? 36	? 4	40
Total	90	10	100

→ June 16

Test statistic (X^2)

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$. O_{ij} : observed, E_{ij} : expected , (row, column) (i, j)

O_{11}	O_{12}
O_{21}	O_{22}

$X^2 \sim \chi^2_{df}$

$df = (r-1)(c-1)$

* each expected count ≥ 5

reject H_0 : $X^2 > \chi^2_{(r-1)(c-1)}$

Do not reject H_0 : $X^2 \leq \chi^2_{(r-1)(c-1)}$

χ^2_{df} p-value $< \alpha$, reject H_0 ; $\geq \alpha$, do not reject H_0

Example

A survey involving 1772 adults gave the following data on alcohol consumption patterns by marital status.

	Drinks per month			
	Abstain	1–60	Over 60	Total
Single	67	213	74	354
Married	411	633	129	1173
Widowed	85	51	7	143
Divorced	27	60	15	102
Total	590	957	225	1772

$df = (4-1)(3-1) = 6$, $X^2 = 94.269$. $\alpha = 0.05$

$\chi^2_6 = 12.592$

$\because X^2 > \chi^2_6$ $\therefore$ associated at 5% level

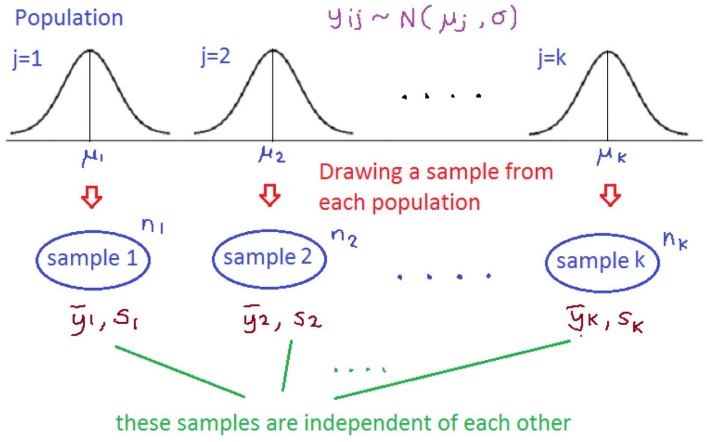
Analysis of variance (ANOVA)

- Comparing 2+ population mean
  independent
- Difference between center of the groups
  the spread within the groups

F-test

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
  H_A : not all μ are equal
  (at least two are different)

→ June 16



$$N = \sum n$$

$$\bar{\bar{y}} = \frac{\sum y_{ij}}{N}$$

The parameters include $\mu_1, \mu_2, \cdots, \mu_k$ and σ.

- Assumptions of the ANOVA:
  1. The k samples drawn from the k populations must be independent of each other.
  2. Within each sample j, the individual observations y_{ij}'s are randomly chosen from the population j. (y_{ij}'s are independent.)
  3. Within each population j, y_{ij}'s follow the Normal distribution with mean μ_j and standard deviation σ. Notice that the k populations have a common standard deviation σ.

(1) variation between groups
- SS_T : treatment sum of squares

(2) variation within groups
- SS_E : error sum of squares

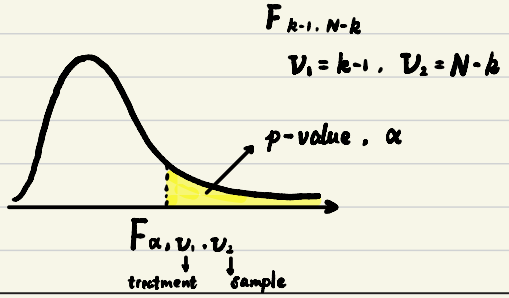
→ $SS_{Total} = SS_T + SS_E$
(total variation of data)

(σ^2)

reject H_0 :
$F > F_{\alpha, v_1, v_2}$, $p < \alpha$

do not reject H_0 :
$F \leq F_{\alpha, v_1, v_2}$, $p \geq \alpha$

F distribution



- Sum of Squares: $SS_{Total} = SS_T + SS_E$

  See Appendix for formulas of the sum of squares.

- Degrees of Freedom (df): $(N-1) = (k-1) + (N-k)$

- The ANOVA table:

Source of Variation	df	Sum of Squares	Mean Squares	F-ratio
Treatments	$k-1$	SS_T	$MS_T = \frac{SS_T}{k-1}$	$\frac{MS_T}{MS_E}$
Error	$N-k$	SS_E	$MS_E = \frac{SS_E}{N-k}$	
Total	$N-1$	SS_{Total}		

Mean Squares (MS) $= \frac{\text{Sum of Squares}}{\text{degrees of freedom}}$