Stats

- Descriptive: collect, summarize, present & analyse data
  - tables, graph, freq. distribution
  - measures of central tendancy
  - measures of dispersion
- Inferential: collect from small group to draw conclusion on large group.
  - Confidence intervals
  - Hypothesis testing

Vocab

- Variable: characteristic of item (or individual)
- Data: ~~Eg~~ different values associated to a variable

- Population: all items for which conclusions are to be drawn
- Sample: Subset of a population
- Inferance: Conclusion on a population drawn from a sample

- Parameter: Measure that describes a characteristic of a population
- Statistic: " " of a sample

Types of variables

- Qualitative (categorical): Value is boolean (Y/N).
- Quantitative: value represents a quantity & can be
  - Discrete
  - Continuous.

Measurement scale

- Nominal: Only descriptive, # designate a class, no implied ranking
- Ordinal: #s represent a scale, but Δ/ratio not measured (ex: 1=good 5=bad)
- Interval: Δ has a meaning, but no common Ø. (ex: 20° ≠ 2×10°)
- Ratio: Δ has a meaning & common zero.

Stat

NP01

Statk method

1. Define variables
2. Collect data (internal or external)
3. Organize (build tables)
4. Visualize (build charts)
5. Analyze (draw conclusions from tables & charts

2.2 Organizing data

Summary table: presents response as frequency or %

Category	freq	%
c_1	x_1	$\%_1$

Contingency table: cross table with 2+ variables to draw relationship conclusions

	x	y	z	total
A	A_x	A_y	A_z	$\sum_x^z A$
B	B_x	B_y	B_z	ΣB
Total	Σx	Σy	Σz	

Array: Ordered #s from smallest to largest

Frequency distribution: Summerizes data by classes. Classes are groups that represents a range of values.

Rules for constructing a frequency distribution

- # of classes btw 5 & 15.
- Determin class intervals by computing range
- Keep class intervals equal
- No overlapping classes.
- No open ended classes if possible
- No null classes if possible

Stats NP02

Class interval = range of a class

Graphing frequency Distributions.

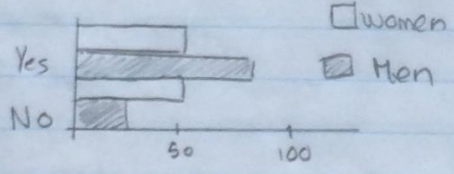
- Histograph: bar chart
- Polygon: regular graph; points are in the middle of F.D.
- Ogive: Cumulative percentage polygon

Pareto Diagram: Combination of ordered histograph + ogive

- Histograph in decending order
- Contains a cuml % line

P45 Pareto principle: majority of occurences happen number of categories few remaining occurences are spread across many categories has vital few & trivial many

Side by side chart: Shows joint response from 2 categorical variables.



Stem & leaf display: helps show distribution & Spikes.
Data is split into groups (stems; rows) and values of groups (leaves) branch out to the right

to construct ex p54

① Sort
② Distribute

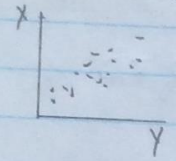
Array: 9.1 9.4 9.7 10.0 10.2 10.2 10.3 10.8 10.11
11.2 11.5 11.6 11.7 11.7 12.2 12.3 12.4 12.8 12.9
13 13.2

Stems	leaves
9	1 4 7
10	0 2 2 3 8 11
11	2 5 6 7 7
12	2 3 4 8 9
13	0 2

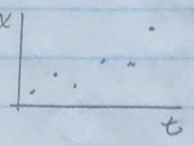
desire normal distribution

Stats NP 03.

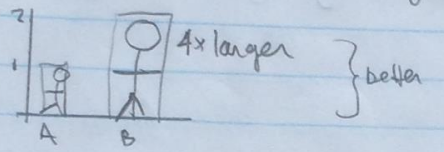
Scatter Diagram



time series: plotted over time



Bad practice: use pictograms that scale



Chapter 3 - Numerical Descriptive Measures.

Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Population mean (parameter): $\mu = \frac{\sum_{i=1}^{N} x_n}{N}$

Median: middle value

Mode: most common value.

Geometric Mean: $\bar{X}_G = \sqrt[n]{\prod_{i=1} x_i}$

Geometric Mean Rate of Return: $\bar{R}_G = \sqrt[n]{\Pi(1+R_n)} - 1$

Sample variance: $S^2 = \frac{\sum (x-\bar{x})^2}{n-1}$ Pop Var: $\sigma^2 = \frac{\sum (x-\bar{x})^2}{N}$

Sample Std. dev $S = \sqrt{S^2}$ Pop Std. dev: $\sigma = \sqrt{\sigma^2}$

Coefficient of Variation $CV = \left(\frac{S}{\bar{x}}\right) 100\%$

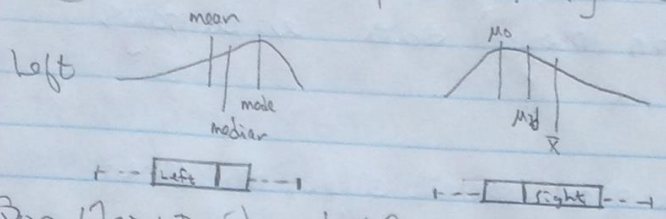
Z scores: How many Std. dev above/below mean

Stats
Npo4. $Z_n = \frac{x_n - \bar{x}}{S}$. Outlier if $|Z_n| > 3$.

Shape

mean < median: negative/left skewed
mean = median: Symetric
mean > median: positive/Right skewed

Not systematic, if few extreme values, skewed that way.



* For discrete, find x s.t. $p(x<X) = 50\%$ for Median

Quartiles: Decides distribution in quarters.

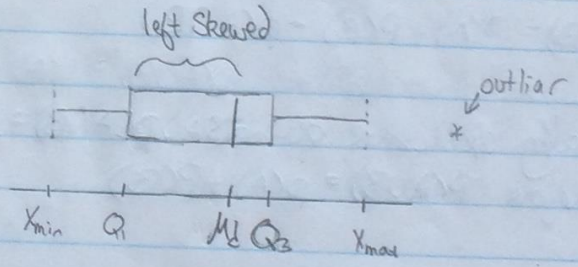
$Q_i = \frac{i(n+1)}{4}$ for $i = 1,2,3$.

- if $Q_i \in \mathbb{N}$, $Q_i = Q_i$
- if exact middle take avg (ex: 2,5 take 2,83)
- Otherwise round to nearest

Inter Quartile Range (IQR) = $Q_3 - Q_1$

5 # Summary
Smallest
Q_1
Q_2 = median
Q_3
Largest

box plot



Outlier

- $X > Q_3 + 1,5\,IQR$ or if $|z_n| > 3$.
- $X < Q_1 - 1,5\,IQR$

Imperial Rule (normal dist)

dist	% pop
$\pm 1\sigma$	68,26
$\pm 2\sigma$	95,4
$\pm 3\sigma$	99,7

Chebycher: regardless of distribution, min pop. within k Std dev is

$pop > \left(1 - \frac{1}{k^2}\right) \times 100\%$

# σs	% pop min
1	0
2	25%
3	89%

Stats
Np05.

CH4- Probability

Probability: # that represents the likelyhood that an event occurs.
- A priori: probability based on the prior knowledge of an event
  ex: cards, dice, coin
- Empirical: probability based on observed data ex: surveys
- Subjective: based on WAG

Exam *
Independant events: Outcome of one does not affect the other
IFF $P(A|B) = P(A)$
Note: Since $P(A \cap B) = P(A|B)\,P(B)$
Independent events imply $P(A \cap B) = P(A)\,P(B)$

Dependant events: Outcome of one influences the other.
If $P(A|B) \neq P(A)$
or $P(A \cap B) \neq P(A) \times P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Mutually exclusive events: $P(A \cap B) = \emptyset$.
therefore $P(A \cup B) = P(A) + P(B)$

Marginal probability
$P(A) = P(A \cap B_1) + P(A \cap B_2) \dots + P(A \cap B_k)$.
where B_k are mutually exclusive & collectively exhaustive.

Collectively exhaustive events: one of the events must occur

Conditional Probability
Since $P(A \cap B) = P(A|B)\,P(B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Stats
NP06.

Contingency table

	M	F	
Black	10	20	30
White	10	10	20
	20	30	50

Joint probability table

	M	F	
Black	0.2	0.4	0.6
White	0.2	0.2	0.4
	0.4	0.6	

Joint probabilities

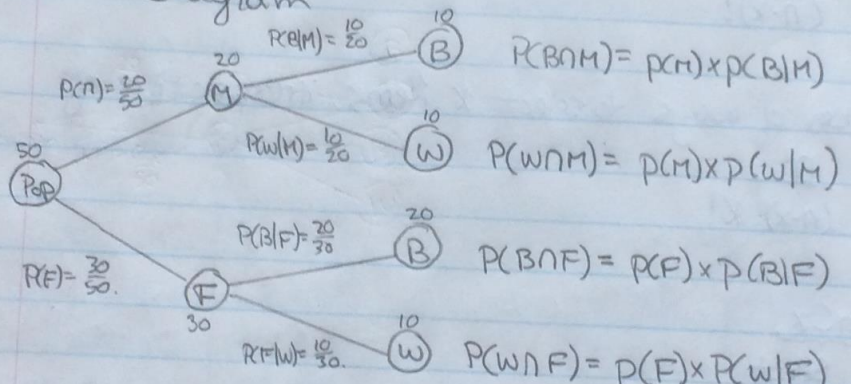
marginal probabilities

Ex: If randomly select F, P(B)=?

$P(B|F)P(F) = P(B \cap F)$.

$P(B|F) = \frac{P(B \cap F)}{P(F)} = \frac{0.4}{0.6} = \frac{2}{3}$.

Tree Diagram



$P(B \cap M) = P(M) \times P(B|M)$

$P(W \cap M) = P(M) \times P(W|M)$

$P(B \cap F) = P(F) \times P(B|F)$

$P(W \cap F) = P(F) \times P(W|F)$.

Bayes' theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Counting

① If k events can occur in n trials, then total possible outcomes $= k^n$

② If k_i event happens on the i^{th} trial, then total possible outcomes $= k_1 \times k_2 \cdots k_n$

③ Number of ways n items can be arranged in order where A, B, C are duplicates: $\frac{n!}{A!B!C!}$

ex MISSISSIPPI $\frac{11!}{4!4!2!}$

S → 4!, I → 4!, P → 2!

④ Permutation: number of ways of arranging x objects selected from n objects in order

$$ {}_nP_x = \frac{n!}{(n-x)!} $$

⑤ Combination: Number of ways to select x items, irrespective of order

$$ {}_nC_x = \frac{n!}{(n-x)!\,x!} $$

$$ \underbrace{{}_nP_x}_{\text{Select \& Order}} = \underbrace{x!}_{\text{order}}\ \underbrace{{}_nC_x}_{\text{Select}} $$

5.1- Discrete

Probability distribution of a discrete random variable is a mutually exclusive list of all the possible numerical outcomes & their probability of occurance.

Expected value $\mu = E(x) = \sum_{i=1}^{n} x_i P(x_i)$

Variance $\sigma^2 = \sum_{i=1}^{n} [x_i - E(x)]^2 P(x_i)$

Std. dev $\sigma = \sqrt{\sigma^2}$

Median when $P(x>X) = 50\%$

Solving

x_i	$P(x_i)$	$x_i P(x_i)$	$[x - E(x)]$	$[x - E(x)]^2$	$[x - E(x)]^2 P(x)$
	0,25				
	025				
	Σ	μ_x / $E(x)$			$\Sigma = \sigma^2$

Binomial Distribution: only 2 possible outcomes.

Probability of x positive outcomes out of n trials given
$p(\text{success}) = p$ & $p(\text{failure}) = 1-p = q$

$$P(X=x|n,p) = {}_xC_n\, p^x q^{n-x}$$

* ${}_nC_x = \frac{n!}{x!(n-x)!}$

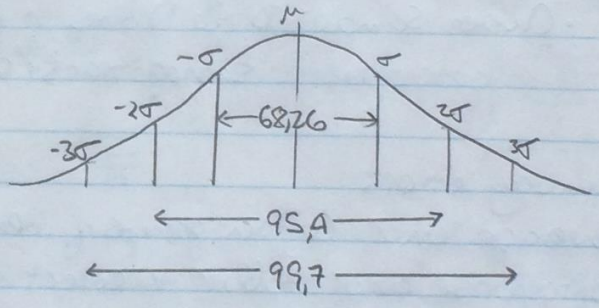
Mean $\mu_B = E(x) = np$.

Variance $\sigma^2 = npq$

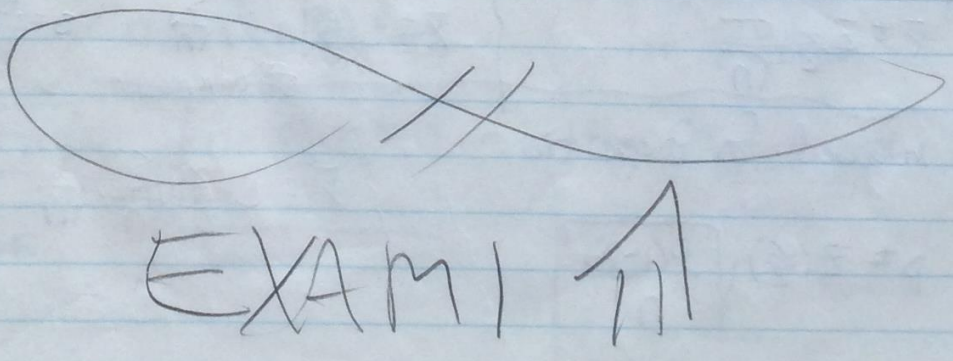
Std. dev $\sigma = \sqrt{\sigma} = \sqrt{npq}$

6. Normal (Gaussian) distribution

x	$P(x)$
$<\mu \pm 1\sigma$	68,26%
$<\mu \pm 2\sigma$	95,4%
$<\mu \pm 3\sigma$	99,7%



$$\boxed{Z = \frac{X-\mu}{\sigma}} \qquad X = Z\sigma + \mu.$$



Stats
Np09

Chap 7 - Sampling Distribution

Frame: listing of all items in a population

Samples

Nonprobability samples: Select items w/o knowing their prob. of selection
- Judgment samples: Ask SMEs their opinion on a subject
- Convenient samples: Items selected are inexpensive (ie. first person that walks by).

Probability Samples: Select items based on prob.
- Simple random sample: every item in the frame has equal chance of being selected.
- Stratified random sample: Subdivide the list of N items in the frame into strata, based on a common characteristic (gender, age, etc...)
- Cluster Sample: Divide N items in frame into clusters, typically geographical segregations (country, election district, household).

Survey errors
- Coverage error: Certain groups of items are excluded from the frame
- Nonresponse error: failure to collect data from from all in sample
- Sampling error: Reflects chance variation from sample to sample.
- Measurement error: Ambigeous wording questions; Hawthorne effect: responder wants to please interviewer; respondent error: over/under zealous respondant

Conf. Interval σ Known

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

σ unknown

$$\bar{X} \pm t_{(\frac{\alpha}{2} | n-1)} \frac{s}{\sqrt{n}} \qquad s = \frac{\Sigma (x-\bar{x})^2}{n-1}$$

Conf interval for Proportion

$$p \pm Z(\tfrac{\alpha}{2}) \sqrt{\frac{P(1-P)}{n}}$$

Sampling dist. of mean: take all possible samples in a population

Central limit theorem: as the sample size gets large, the sample distribution of the mean will be normal, regardless of shape of distribution of individuals

	Mean	Var.	Std. dev	Size	Z	Proportion
Sample	$\bar{X}$	$s^2=\frac{\Sigma(x-\bar{x})^2}{n-1}$	s	n	$\frac{x-\bar{x}}{s}$	p
Population	μ	$\sigma^2=\frac{\Sigma(x-\mu)^2}{N}$	σ	N	$\frac{x-\mu}{\sigma}$	π
Sampling distribution	$\bar{\bar{X}}$	$\sigma_{\bar{x}}^2$	$\sigma_{\bar{x}}$	n_s	$\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$	

Population Mean $\mu=\frac{\sum X_i}{N}$ or from samples $\mu=\frac{\sum \bar{X}}{n_s}$

*95.4% of all sample's means will fall $\pm 2\sigma$ of $\bar{X}$

Population Std. dev $\sigma=\sqrt{\frac{\Sigma(X_i-\mu)^2}{N}}$

Sample Std dev $s=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

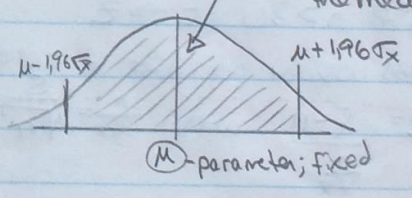
Standard Error of mean $\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$

Z for sampling distribution of mean

$$Z=\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\bar{X}=\mu+Z\frac{\sigma}{\sqrt{n}}$$

95% of ranges $\bar{X}\pm 1.96\sigma_{\bar{x}}$ contain the mean



a) Compute Sample size to ensure % confidence.

$Z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}$ $\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$

$\frac{Z\sigma}{\sqrt{n}}=\bar{X}-\mu$ ← Error E

$$n=\frac{Z^2\sigma^2}{\epsilon^2}$$

for proportion

$$n=\frac{Z^2(\frac{\alpha}{2})\pi(1-\pi)}{\epsilon^2}$$

Stat
N pll

Sampling Distribution of proportion

Population proportion: π is the proportion of all items in the population with a characteristic of interest (a parameter)

Sample proportion: p is the sample proportion (a statistic) that is used to estimate π

$$p = \frac{X}{n} = \frac{\#\text{ w/ characteristic}}{\text{Sample size}}.$$

Standard error of proportion

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad z = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

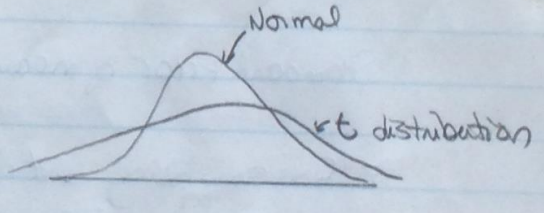
Student t distribution

When to use

① Do not know σ & use s to estimate σ

② If $n > 30$; $t \to z$.

③ Must assume pop. dist is normal



$t = \frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{x}}}$ where $\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$ instead of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Degrees of freedom (d.f.): $n-1$

Degree of Confidence: $1-\alpha$

$$P(\bar{X} - t\hat{\sigma}_{\bar{x}} \le \mu \le \bar{X} + t\hat{\sigma}_{\bar{x}}) = 1-\alpha$$

—FINAL—

Wording:
Probability that the interval $[\bar{X} - t\hat{\sigma}_{\bar{x}}, \bar{X} + t\hat{\sigma}_{\bar{x}}]$ contains the population mean is $1-\alpha$.

$$s = \sqrt{\frac{\Sigma(\bar{X}-X)^2}{n-1}} \qquad \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$t_{(n-1),(1-\alpha)} = \frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{x}}}$$

$(n-1)$: degree of freedom; $(1-\alpha)$: confidence %

Stat Np12.

Ex 8.13 p292

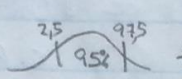
A	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-3,5	12,25
1	-3,5	
1	-3,5	
1	-3,5	
8	3,5	
8	3,5	
8	3,5	
8	3,5	
Σ 36		Σ 98
$\bar{x}$ 4,5		$\bar{x}$=4,5

B	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-3,5	12,25
2	-2,5	6,25
3	-1,5	2,25
4	-0,5	0,25
5	0,5	0,25
6	1,5	2,25
7	2,5	6,25
8	3,5	12,25
		Σ 42

Find confidence intervals for 95% certainty

$S_A = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = 3,74$ $\quad S_B = \sqrt{\frac{\Sigma(x-\bar{x})}{n-1}} = 2,45$

$\hat{\sigma}_{\bar{x}_A} = \frac{S_A}{\sqrt{n}} = 1,32$ $\quad \hat{\sigma}_{\bar{x}_B} = \frac{S_B}{\sqrt{n}} = 0,87$



$t_{7,25} = 2,3646$ $\quad t_{7,975} = 2,3646$.

Want $P(\bar{x} - t_{n-1,1-\alpha}\hat{\sigma}_{\bar{x}} \leq \mu \leq \bar{x} + t_{n-1,1-\alpha}\hat{\sigma}_{\bar{x}}) = 1-\alpha$.

A: $P(4,5-(2,36)(1,32) \leq \mu \leq 4,5+(2,36)(1,32)) = 0.95$.

$P(1,36 \leq \mu \leq 7,64) = 0,95$.

B: $p(2,45 \leq \mu \leq 4,55) = 0,95$.

Confidence interval for Proportion

*do not use t for proportion

$P(p - z\sigma_p \leq \pi \leq p + z\sigma_p) = 1-\alpha$ $\quad z = \frac{p-\pi}{\sigma_p}$ $\quad \sigma_p = \sqrt{\frac{p(1-p)}{n}}$

To build a $1-\alpha$ confidence interval for population

① Given $x, n \Rightarrow$ calculate $p = x/n$.

② Calculate σ_p

③ find $z_{1-\alpha}$

Determining Sample size of Proportion

$\sigma = \sqrt{\frac{p(1-p)}{n}}$ or $\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}$ $\quad z = \frac{p-\pi}{\sigma} \Rightarrow \boxed{\sigma = \frac{p-\pi}{z}}$

$n = \frac{\pi(1-\pi)}{\sigma^2} \rightarrow \boxed{n = \frac{z^2\pi(1-\pi)}{(p-\pi)^2}}$ when $(p-\pi)^2 = E^2$.

Use π if have educated guess.

Otherwise use $n = \frac{z^2 p(1-p)}{(p-\pi)^2}$

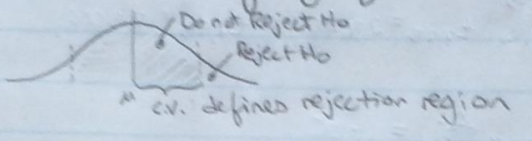
Otherwise, use $P = 0,5$ to yield largest estimate of n to ensure contingency

$$n = \frac{z^2(0,25)}{E^2} = 0.25\left(\frac{z}{E}\right)^2$$

Hypothesis testing (p326).

Critical value

1) State H_0: null and H_1: Alternative
2) State the level of significance α & Critical value (C.V.)

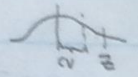


α: willingness to have a rejection when true

3) Perform the test
   a) from α; determin C.V. in terms of Z (ex: 95% certainty α=0,05 C.V.=1,96).
   b) find $Z_{stat} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

* if σ unknown & small sample use $t_{stat,\alpha} = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}$ where $\sigma_{\bar{x}} = \frac{S}{\sqrt{n}}$ (S: sample std dev)

* for proportion, always use Z. (need n large enough)

4) Draw conclusion
   if $|Z_{stat}| > |C.V.|$ then  Reject Ho.

otherwise, if Z is within C.V.; do not Reject

5) Interpret Conclusion

ex
Claim $\mu = 35$
Test: $\bar{X} = 32$; $n = 100$; $\sigma = 15$.
Set: $\alpha = 0,05$.
1) H_0: $\mu = 35$ H_1: $\mu \neq 35$.
2) for $\alpha = 0,05$; C.V. = 1,96
3) $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{32-35}{15/\sqrt{100}} = -2$
4) Since $|Z| > |C.V.|$ Reject Ho.

* if α were 0,01; C.V. = 2,575.
then $|Z| < |C.V.|$ Do not reject

ex 9.25 p342

Claim $\mu = 8,17$.
Test: $n = 50$; $\bar{X} = 8,159$; $S = 0,051$
Set $\alpha = 0,05$.
1) H_0: $\mu = 8,17$ H_1: $\mu \neq 8,17$.
2) for $\alpha = 0,05$; C.V. = 1,96.
3) $t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{8,159-8,17}{0,051/\sqrt{50}} = -1,5251$
4) Since $|t| < |C.V.|$ do not Reject

$$Z_{STAT} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

$$Z_{stat} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Stat 1
p14

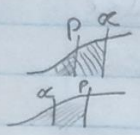
P333 P-Value: probability of getting a test statistic equal or more extreme than the sample result, assuming H_0 is true.

Steps

1) State H_0: $\bar{x}=\mu$ & H_1: $\bar{x}\neq\mu$
2) State α
3) Compute p-value

$z_{stat} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ then p-Value $= [1-p(z_{stat})] + p(-z_{stat})$. if z, do P(z) if t, Reverse lookup. *2 if 2 tail test

4) Draw conclusion



if p-Value $< \alpha$ Reject H_0.

if p-Value $> \alpha$ Do not Reject H_0

One side test: H_1 is what you want to prove (increase or decrease).
z/t(α) p.V. Do. not *2.

2 side test:
z/t($\frac{\alpha}{2}$) p.V. is p(tstat) *2

ex) Claim $\mu = 8,17$.

Test $n=50$; $\bar{x}=8,159$; $S=8,17$.

Set $\alpha = 0,05$

1) H_0: $\mu = 8,17$.
2) $\alpha = 0,05$; C.V. = 1,96.
3) $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = -1,525$.

P-Value = $p(t \geq |1,525|)$.

Reverse lookup: t is between 0,1 & 0,05.

2 tail test: p-Value is between 0,1 & 0,2.

4) either way, p-Value $> \alpha$
⇒ do not reject H_0.

When doing proportions, use z, not t.
⇒ make sure n is large enough.

P-Value meaning: the probability of getting $t_{stat} >$ [C.V.] or $t_{stat} <$ [-C.V.] given that H_0 is true is [p-Value].

$z_{stat} >$ C.V.	Reject	p-V $< \alpha$
$z_{stat} <$ C.V.	Do not Reject	P.V. $> \alpha$

Stats
N p15.

ERRORS. P328

Type I: Reject H_0 when it is actually true. $P(\text{type I}) = \alpha$.

Type II: Did not reject H_0 when it's actually false (missed opportunity to take corrective action) $p(\text{type II}) = \beta$.

α: level of significance $\alpha = p(\text{type I})$

β: Risk of committing type II $\beta = p(\text{type II})$.

Confidence coefficient $(1-\alpha)$: prob. do not reject H_0 when it should.

Power of a stat test $(1-\beta)$: prob to reject when actually false

	H_0 is true	H_0 is false
Do not Reject H_0	No ERROR	Type II ERROR
	Confidence $1-\alpha$	Risk β.
Reject H_0.	Type I ERROR	No ERROR
	significance α	power $1-\beta$

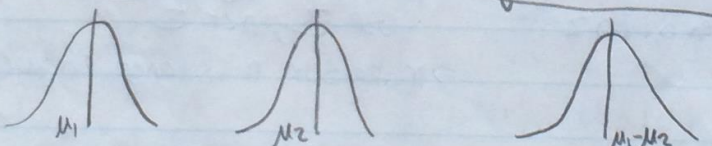
Fix α; for $n\uparrow \Rightarrow P(\text{type II})\downarrow$

If bad consequence of making type I, make $\alpha\downarrow$

Never use t on proportions $\Rightarrow$ Always Z.

for $t \Rightarrow$ Assume normal distribution of population.

Sampling Distribution of the distance btw means.



$\sigma_{\bar{x}_1} = \frac{\sigma_1}{n_1}$ $\sigma_{\bar{x}_2} = \frac{\sigma_2}{n_2}$ $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ ← population

Since most times σ_{pop} is not available, use t.

Pooled Variance t-test

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

*Assume population is normally dist.

exp 375 10.9

① $H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

② d.f. $= (n_1-1)+(n_2-1) = 38$.
$\alpha = 0{,}05$; C.V. $= t(38, \frac{\alpha}{2}) = 2{,}0244$

③ $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} = 0{,}356$.

④ $[t = 0{,}356] < [\text{C.V.} = 2{,}02]$
⇒ Do not Reject H_0.

⑤ Based on step ② criteria, it is possible that both samples come from pops. w/ same mean

Confidence Interval

$$P\left[(\bar{X}_1 - \bar{X}_2) - t(\tfrac{\alpha}{2}, n_1+n_2-2)\sqrt{S_p^2\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 + \bar{X}_2) + t(\tfrac{\alpha}{2}, n_1+n_2-2)\sqrt{S_p^2\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)}\right] = 1-\alpha$$

Paired t-test: Comparing means of 2 related pops.

n	X_A	X_B	$D = X_A - X_B$
1	90	85	-5
2	95	88	7
3	87	87	0
4	55	86	-1
5	90	82	8
6	91	82	12
7	85	70	15
8	88	72	16
9	92	80	12
		Σ	74

$\bar{D} = 8{,}2$.

① $H_0: \mu_D = \emptyset$
$H_1: \mu_D \neq 1$

② $\alpha = 0{,}05$ d.f. $= 8$.
$t(\frac{\alpha}{2}, 8) = 2{,}306$.

③ $S_p = \sqrt{\frac{\Sigma(D-\bar{D})^2}{n-1}} = 6{,}1$

$t = \frac{8{,}2 - \emptyset}{6{,}1/\sqrt{9}} = 4{,}02$

④ $t = 4{,}02 > \text{C.V.} = 2{,}306$.
Reject H_0.

⑤ Based upon criteria in ②, it is unlikely that this sample came from populations where diff. $= \emptyset$.

Confidence Interval of the mean Difference

$$P\left[\bar{D} - t(\tfrac{\alpha}{2}, n-1)\frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + t(\tfrac{\alpha}{2}, n-1)\frac{S_D}{\sqrt{n}}\right] = 1-\alpha.$$

⇒ Here $P(3{,}5 \leq \mu_D \leq 12{,}9) = 0{,}95$
conf. interval does not contain $H_0 \Rightarrow \emptyset$.

Stat NPT

Z-test for Diff. proportions

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \quad \text{where} \quad \bar{p} = \frac{\bar{X}_1 + \bar{X}_2}{n_1 + n_2} \Big\} \text{pooled estimate of pop. proportion}$$

or

p391/10.29 A) $P_1 = \frac{X_1}{n_1} = 0{,}567$ ① $H_0: \pi_1 = \pi_2$ ③ $z = \frac{\Delta p - \Delta \pi}{\sqrt{\bar{p}\bar{p}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -7{,}35$ ⑤ Based upon criteria established in 2 it is unlikely that these 2 samples came from pops. w/ same proportion

$P_2 = \frac{X_2}{n_2} = 0{,}862$ $H_1: \pi_1 \neq \pi_2$.

$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = 0{,}72$ ② $\alpha = 0{,}01$ ④ Since $z = 7{,}35 > C.V. = 2{,}575$

$C.V. = z(\frac{\alpha}{2}) = 2{,}575$. Reject H_0

B) P-Value. $P(Z < -7{,}35 \mid \pi_1 - \pi_2) + P(Z > 7{,}35 \mid \pi_1 - \pi_2) \simeq \emptyset$.

If H_0 is true, the probability of getting 2 samples w/ these proportions $\simeq \emptyset$.

Confidence Interval for $\Delta\pi$

$$\text{⑥ } P\left[(p_1 - p_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq \pi_1 - \pi_2 \leq (p_1 - p_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right] = 1 - \alpha$$

do not assume $P_1 = P_2$ since rejected H_0

$$P[-0{,}3934 \leq \pi_1 - \pi_2 \leq -0{,}1965] = 0{,}95$$

CH 11 p 418. One Way ANOVA F-test

Compares Among (A) variation & Within (W) variation

Assumptions

① Populations are normally distributed.

② Populations have equal variances (Homo sceda sticity)

③ Samples are randomly & Independantly drawn

Total Variation (SST) = Among group Variance (SSA) + Within Group Var. (SSW)

d.f. = n-1 d.f. = c-1 d.f. = n-c

If Among > Within, probably came from dif. populations.

$$F_{STAT} = \frac{\text{Among}}{\text{Within}}.$$

Stat N p18

ANOVA

① H_0: $\mu_1 = \mu_2 = \mu_3$.
H_1: Not all means are equal

② d.f. A = (C) + 1 ← #groups
d.f. W = $n_{tot} - C$
C.V. = $F(\alpha, \frac{d.f.A}{d.f.w})$ @p 803

$\alpha = 0,05$.
d.f. A = 3-1 = 2
d.f. W = 9-3 = 6
C.V. = $F(0,05 | \frac{2}{6}) = 5,14$.

③ $\bar{\bar{X}} = \frac{\sum \text{all } X}{n_{total}}$

$\bar{\bar{X}} = \frac{60+120+100}{9} = 31,3$

SSW = $\sum$ all variances

SSW = 200+200+200 = 600

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$
10	-10	100
20	∅	∅
30	+10	100
Σ 60		200
$\bar{X}_1$ 20		

X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
20	-10	100
30	∅	∅
30	∅	∅
40	10	100
Σ 120		200
$\bar{X}_2 = 30$		

X_3	$X_3 - \bar{X}_3$	$(X_3 - \bar{X}_3)^2$
40	-10	100
60	+10	100
Σ 100		200
$\bar{X}_3 = 50$		

MEAN SQUARE WITHIN

$MSW = \frac{SSW}{n_t - C} = \frac{600}{9-3} = 100$

SSA = Compare $\bar{X}_i$ w/ $\bar{\bar{X}}$

$SSA = n_1(\bar{X}_1 - \bar{\bar{X}}) + n_2(\bar{X}_2 - \bar{\bar{X}}) + n_3(\bar{X}_3 - \bar{\bar{X}})$

SSA = 1088,89.

n_i	$\bar{X}_i$	$\bar{X}_i - \bar{\bar{X}}$	$(\bar{X}_i - \bar{\bar{X}})^2$	$n_i(\bar{X}_i - \bar{\bar{X}})^2$
3	20	-11,1	123,21	369,63
4	30	-1,1	1,21	4,84
2	50	18,9	357,21	714,42
			Σ	1088,89

MEAN SQUARE AMONG.

$MSA = \frac{SSA}{C-1} = \frac{1088,89}{3-1} = 544,45$.

Source	d.f.	sum of Squares	Variance (Mean square)	F
Among	C-1	SSA	$MSA = \frac{SSA}{d.f.a}$	$F_{stat} = \frac{MSA}{MSW}$
Within	$n_t - C$	SSW	$MSB = \frac{SSB}{d.f.b}$	
Total	$n_t - 1$	SST		

useless.

$F_{stat} = \frac{MSA}{MSW} = \frac{544,45}{100} = 5,44$.

④ $F_{stat} = 5,44 > F = 5,14$.
Reject H_0.

⑤ Based on criteria established in 2, it is improbable that the samples came from populations w/ same mean.

Stat
N p19.

CHI2 - Chi SQUARE

Difference between 2 proportions.

$\psi^2_{STAT} = \sum \frac{(f_o - f_e)^2}{f_e}$ f_o = frequency observed

f_e = frequency expected (Assume Y/N of population is true, apply on categories by scaling)

$d.f. = (\#Rows - 1)(\#Cl - 1)$

$C.V. = \psi^2(\alpha, d.f.)$ p. 802.

ex 12 p 473

① $H_0: \pi_1 = \pi_2$

$H_1: \pi_1 \neq \pi_2$

② $\alpha = 0,01$

$d.f. = (2-1)(2-1) = 1$

$C.V. = \psi^2(0,01/1) = 6,635$

	MEN		WOMEN		
	f_o	f_e	f_o	f_e	
Y	136	$\frac{360}{500} \times 240 = 172,8$	224	$\frac{360}{500} \times 260 = 187,2$	360
N	104	$\frac{140}{500} \times 240 = 67,2$	36	$\frac{140}{500} \times 260 = 72,8$	140
	240	check 240 ✓	260	260 ✓	500

③

Cell	f_o	f_e	$f_o - f_e$	sq	$\frac{(f_o - f_e)^2}{f_e}$
YM	136	172,8	-36,8	1354,24	7,837
NM	104	67,2	36,8	1354,24	20,1524
YW	224	187,2	36,8	1354,24	7,2342
NW	36	72,8	-36,8	1354,24	18,6027
				Σ	53,8258

④ $\psi^2_{stat} = 53,8258 > \psi^2_{max} = 6,635$

Reject H_0

For p-Value, reverse lookup on ψ^2 table.

$\Rightarrow p(\psi^2 > 53,82)$

① 1 d.f., largest ψ^2 is $p(\psi^2 > 7,879) = 0,005$.

therefore $p(\psi^2 > 53,82) < 0,005$. is the only conclusion possible

Stat

Np20

Chi Square test for 2+ proportions.

p481 ex13,19 ① $H_0: \pi_1 = \pi_2 = \pi_3$

H_1: Not all equal

② $\alpha = 0,05$.

$df. = (R-1)(C-1) = (3-1)(2-1) = 2$

$\chi^2(0,05/2) = 5,991$.

③ $\chi^2_{stat} = \sum_{cells} \frac{(f_o - f_e)^2}{f_e} = 5,311$

④ $\chi^2_{stat} = 5,311 < \chi^2 = 5,991$

Do not reject.

	16-19		30-49		50-64		
	F_o	F_e	F_o	F_e	F_o	F_e	total
Y	50	$\frac{154}{600} \times 200 = 51,33$	42	$\frac{154}{600} \times 200 = 51,33$	62	$\frac{154}{600} \times 200 = 51,33$	154
N	150	$\frac{446}{600} \times 200 = 148,67$	158	$\frac{446}{600} \times 200 = 148,67$	138	$\frac{446}{600} \times 200 = 148,67$	446
	200		200		200		600

p-Value $\Rightarrow$ 2 d.f. $\chi^2_{stat} = 5,311$

$p = [0,1 - 0,05]$.

Between 5% & 10%.

Chi Square for Independance

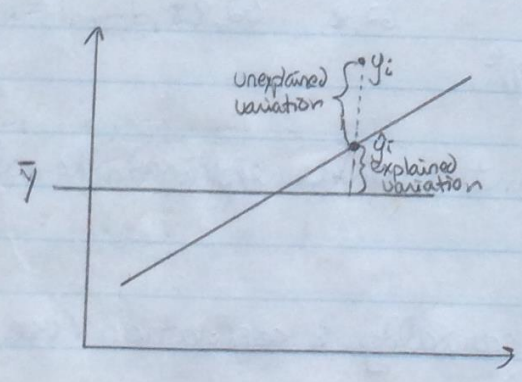
Ho: There is no relationship between "Gender" & "Shopping habits"

$\Rightarrow f_o \simeq f_e$; $\chi^2_{stat} \simeq \emptyset$

d.f. = (# Rows - 1) x (# Columns - 1).

EXAM II

Chapter 13. Simple linear Regression



total = regression + error

$SST = SSR + SSE$

$\sum (y - \bar{y})^2 = \sum (\hat{y}_o - \bar{y})^2 + \sum (y_i - \hat{y}_i)$

total variation = explained variation + unexplained variation

Coefficient of Determination

$R^2 = \frac{\sum_i (\hat{y}_o - \bar{y})^2}{\sum (y - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i (y - \bar{y})^2}$

$\hat{y}_i = b_o + b_i x_i$

Coefficient of correlation $R = \sqrt{R^2}$

Stats Npal

Linear Regression

$\hat{y}_i = b_0 + b_1 x_i$

$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{ssXY}{ssX}$

$ssXy = \sum x \cdot y - \frac{(\sum x)(\sum y)}{n}$

$ssX = \sum x^2 - \frac{(\sum x)^2}{n}$

Quick b_1

$b_1 = \frac{(\sum xy) \cdot n - (\sum x)(\sum y)}{(\sum x^2) n - (\sum x)^2}$

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = \bar{y} - b_1 \cdot \bar{x}$

						for ssXy & ssX		Best match equation	for coefficient of correlation		
X	Y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$X \cdot Y$	X^2	$\hat{y} = b_0 + b_1 X$	$(y-\bar{y})^2$	$(y-\hat{y})^2$	$(\hat{y}-\bar{y})^2$
0	5	-2	-4,8	9,6	4	∅	∅	4,8+2,5(∅)=4,8	23,04	$(0,2)^2$=0,04	$(5)^2$=25
1	7	-1	-2,8	2,8	1	7	1	4,8+2,5(1)=7,3	7,84	$(0,3)^2$=0,09	$(2,5)^2$=6,25
2	10	0	0,2	∅	0	20	4	9,8	0,04	0,04	∅
3	12	1	2,2	2.2	1	36	9	12,3	4,84	0,09	6,25
4	15	2	5,2	10.4	4	60	16	14,8	27,04	0,04	25
Σ 10	49		Σ	25	10	123	30	Σ	62,80 total	0,30 unexplained variation	62,5. explained
$\bar{X}$= 2	9,8										

$ssXy = 123 - \frac{10 \times 49}{5} = 25$

$b_1 = \frac{25}{10} = 2,5.$

$ssX = 30 - \frac{(10)^2}{5} = 10.$

$b_0 = (9,8) - (2,5)(2) = 4,8.$

$\hat{y} = 4,8 + 2,5 x_i$

P535 Other method.

Coefficient of Determination

$ssT_{total} = \sum y^2 - \frac{(\sum y)^2}{n} =$ | $SST = (y-\bar{y})^2$

$ssE_{error} = \sum y^2 - b_0 \sum y - b_1 \sum xy$ | $SSE = (y-\hat{y})^2$

$ssR_{explained} = b_0 \sum y + b_1 \sum xy - \frac{(\sum y)^2}{n}$ | $SSR = \sum (\hat{y}-\bar{y})^2$

$R^2 = \frac{\text{Explained var}}{\text{tot var}} = \frac{ssR}{ssT} = 1 - \frac{\text{unexplained}}{\text{total}} = 1 - \frac{SSE}{SST}$

$R^2 = \frac{62,5}{62,8} = 1 - \frac{0,3}{62,8} = 0,9968$

R = % of the variation in the dependant variable explained the value of the dependant variable

- Supurious correlation: Confusing causality & corrdation (use wrong independant variable)

Stat Np22

13.7 P547 Inferences About the Slope

β_1 = population slope

To test how close b_1 is to β_1

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}}$$

Standard error of estimate p537

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$$ ← error sum of squares. (unexplained variance)

P548 $$S_{b_1} = \frac{S_{yx}}{\sqrt{SSX}}$$

$SSX = \Sigma(x-\bar{x})^2$

if σ_{xy} is known, use z

ex)

① $H_0: \beta_1 = 0$ No relationship btw X & Y.
$H_1: \beta_1 \neq 0$: there is a relationship (use one sided for ± correlation)

② $\alpha = 0.05$ d.f. $= n-2 = 3$.
c.v. $= t(0.05|3) = 3.182$

③ $$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$S_{yx} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} = \sqrt{\frac{0.30}{5-2}} = \sqrt{0.1}$$

$$\sqrt{SSX} = \sqrt{\Sigma(x-\bar{x})^2} = \sqrt{10}$$

$$S_{b_1} = \frac{S_{yx}}{\sqrt{SSX}} = \sqrt{\frac{0.1}{10}} = 0.1.$$

Assume H_0

$$t_{stat} = \frac{2.5 - 0}{0.1} = 25$$

④ $t_{stat} > t_{c.v.}$: Reject H_0.

⑤ There might be a relationship btw X & Y

Confidence interval for population slope β_1

$$b_1 \pm t_{\frac{\alpha}{2}} S_{b_1}$$

↑ $t_{c.v.}$, Not t_{stat}

d.f. $= n =$ # ind Var $- 1$

Stat Np23

F-test for the slope
MSR = explained variance (variance due to regression)
MSE = Error/unexplained variance

① $H_0: \beta_1 = \emptyset$ No relationship btw. X & Y
$H_1: \beta_1 \neq \emptyset$

② $\alpha = 0,05$
d.f. numerator = # of independant variables = 1
d.f. denominator = $n_t - 2 = 5 - 2 = 3$.
C.V. = $F(0,05/1/3) = 10,13$

③ $F_{stat} = \frac{\text{Explained Variation (MSR)} / \text{d.f. num.}}{\text{Error/unexpected variation (MSE)} / \text{d.f. error}} = \frac{62,5/1}{0,3/3}$

④ $F_{stat} > F_{c.v.}$ Reject H_0.

Confidence interval for the slope

$$\beta_1 = b_1 \pm t(\tfrac{\alpha}{2}/n-2) S_{b_1}$$

P551 t test for Correlation Coefficient ρ
is there a statistically significant relationship b.tw. X & Y.

① $H_0: \rho = \emptyset$
$H_1: \rho \neq \emptyset$

② $\alpha = 0,05$
$C.V. = t(\frac{\alpha}{2}/n-2)$

③ $t_{stat} = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$

$r = +\sqrt{r^2}$ if $b_1 > \emptyset$
$r = -\sqrt{r^2}$ if $b_1 < \emptyset$

$$r = \frac{cov(X,Y)}{S_X S_Y}$$

$$Cov(XY) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$$

$$S_X = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n-1}}$$

Start
NP24

P555 Confidence interval for mean value y_i for a given X_i

$$P\left(\hat{y}_i - t\left(\frac{\alpha}{2}\middle|n-2\right) S_{yx}\sqrt{h_i} \le \boxed{\mu_{Y|X=X_i}} \le \hat{y}_i + [\ldots]\right) = 1-\alpha$$

P537 S_{yx} = Standard [error/deviation] of the estimate

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}}$$

$$h_i = \frac{1}{n} + \frac{(X-\bar{X})^2}{SSX} \qquad SSX = \Sigma(X_i - \bar{X})^2$$

$\hat{y}_i = b_0 + b_1 X_i$

$\mu_{Y|X=X_i}$ = mean value of Y when $X = X_i$

P556 Prediction interval for an individual value of y for a given X_i

$$P\left[\hat{y} - t\left(\frac{\alpha}{2}\middle|n-2\right) S_{yx}\sqrt{1+h_i} \le \boxed{Y_{X=X_i}} \le \hat{y} + [\ldots]\right] = 1-\alpha$$

P586 & 563. ANOVA Summary table for Overall F test.

Source	d.f	Sum of Squares	Mean Square (Variance)	F
Regression	k = # ind. var.	SSR	SSR/k	$F_{stat} = \frac{MSR}{MS\bar{E}}$ $\quad d.f. = \frac{k}{n-k-1}$
error	n-k-1	$SS\bar{E}$	SSE/(n-k-1)	
Total	n-1	SST		

H_0: model is shit

H_1: at least 1 ind var is related to dep variable

$SSR = \Sigma(\hat{Y}_i - \bar{Y}_i)^2$ Regression (explained) variation

$SSE = \Sigma(y_i - \hat{y}_i)^2$ Error (unexplained) variation

$SST = \Sigma(y_i - \bar{y}_i)^2$ Total sum of squares.

Coef. of Determination $r^2 = \frac{SSR}{SST}$.

$$r^2_{adj} = 1 - \left[(1-r^2)\frac{n-1}{n-k-1}\right]$$

Stat

Np 25

	Coefficients	Std. Error	t-Stat	P-value
Intercept (b_0)	b_0			
Price (b_1)	b_1			
Promotion (b_2)	b_2			
...				

Most important independent variable is the one w/ highest t

test All variables one by one to see if relevant

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$