STAT 255 - Final Course Material

mercredi 21 avril 2021 10:41

Set 1 - Populations and Samples
Set 2 - Measures of Location and Variability
Set 3 - Probability and Counting
Set 4 - Probability Axioms
Set 5 - Conditional Probability
Set 6 - Mutually Exclusive and Independence
Set 7 - Multiplication Rule and Bayes' Theorem
Set 8 - Probability Density Functions
Set 9 - Expectation
Set 10 - Variance
Set 11 - The Binomial Distribution
Set 12 - The Poisson Distribution
Set 13 - Continuous Densities
Set 14 - The Normal Distribution
Set 15 - Sampling
Set 16 - Expectation and Variance
Set 17 - Inferences About the Population Mean
Set 18 - The T Distribution
Set 19 - Hypothesis Testing
Set 20 - Hypothesis Testing of A Population Mean
Set 21 - Confidence Intervals and Power
Set 22 - Confidence Intervals for Population Proportions
Set 23 - Sample Size and Hypothesis Tests for a Population Proportion
Set 24 - Testing Two Proportions
Set 25 - Comparing Two Population Means
Set 26 - Comparing Two Population Means
Set 27 - Paired Tests

SET 1 - Populations and Samples

☐ Descriptive Stats

Definitions

- Population: a collection of objects
- Sample: a selection from the population
- Parameter: a descriptive measure of the population (μ, σ, σ^2)
- Statistic: a descriptive measure of the sample ($\bar{x}, s, s^2$)
- Random Variable (r.v.):
  - discrete r.v.: set of all possible values are finite or are infinite and countable $(0, 1, 2, \ldots, 100)$.
  - Continuous r.v.: set of all possible values are infinite or are infinite and uncountable $(0, 1)$.

SET 2 - Measures of Location and Variability

☐ Descriptive Stats

Definitions

- Numerical Data: Real Numbers (length, temperature, Scores).
- Categorical Data: Categories (eye colour, customer service survey).
  - Ordinal Data: categories can be Ordered (letter grades)
  - Nominal Data: categories are not ordered (eye colour)

- Total frequency should be 100%.
- Categorical Data is useful for charts (Bar, Pie)
- Numerical Data is useful for chats (Histogram)

- Sample Mean : $\bar{x}$ $\frac{1}{n}\sum_{i=1}^{n} x_i$ (average)
- Sample median : $\tilde{x}$ middle observation of sorted data.
- Mode : common value (not limited to 1 mode).

Note: Don't round numbers.
$\bar{x}$ is sensitive to outliers
$\tilde{x}$ is not sensitive to outliers

- Sample variance : s^2 $\left(\sum_{i=1}^{n} (x_i-\bar{x})^2\right)/(n-1)$
- sample standard deviation : s $\sqrt{\text{variance}}$

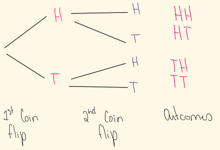
SET 3 - Probability and Counting

☐ Probability

Definitions

- Sample Space : S All possible outcomes of an experiment
  - → Exhaustive (all possible outcomes)
  - → Mutually Exclusive (no overlap between outcomes)
- Simple Event : A single outcome from S.
- Event : A set of one or more simple events. (P(A)).
- Probability :
  - Subjective : Estimated Guess (Expert Opinion)
  - Relative frequency : Estimate the prob. based off of many observation.
    P(A) = f/n = # times events occurred / # experiment
  - ☆ - Classical Method : P(A) = n(A)/n(S) = # of ways A can occur / # of outcomes in S
- Tree Diagrams: used to list all possible outcomes.

ex.



S = { HH, HT, TH, TT }

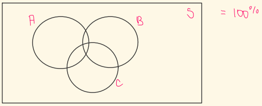
SET 4 - Probability Axioms

☐ Probability

Definitions

- A' : Complement of A
- P(A') = 1 - P(A) & P(A) = 1 - P(A')
- P(S) = 1
- P(A) ≥ 0
- $P(A_1 \text{ or } A_2 \text{ or } A_3) = P(A_1) + P(A_2) + P(A_3)$

- P(A or B) = P(A) + P(B) - P(A and B)
- P(A or B or C) = P(A) + P(B) + P(C) - P(A and B) - P(A and C)
  - P(B and C) + P(A and B and C)



Note: ∘ or → union (inclusive or)
∘ and → intersection

SET 5 - Conditional Probability

☐ Probability

Formula: probability of A given B (i.e. the probability that event A will occur given that we know event B has occured).

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

ex. If we know that A occurs, what is the probability that B has occured.

Given $$p(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

Relative Risk: Ratio of probabilities. It measures the likelihood of having a disease or condition given the subject has a risk factor, versus having the disease or condition given not having the risk factor.

Notation: D = has disease (or Condition), $\mathcal{E}$ = has risk factor

Formula:

$$RR = \frac{P(D|\mathcal{E})}{P(D|\mathcal{E}')}$$

Note: $RR = 1$
$RR < 1$
$RR > 1$

SET 6 - Mutually Exclusive and Independence

☐ Probability

Definition

- Events A and B are independent when ...

  $P(A|B) = P(A)$ (or $P(B|A) = P(B)$)

- Events are independent when...

  $P(A \text{ and } B) = P(A) \cdot P(B)$

Definition

- Events A and B are mutually exclusive if the cannot occur at the same time (i.e. no overlap).

  $P(A \text{ and } B) = 0$

SET 7 - Multiplication Rule and Bayes' Theorem

☐ Probability

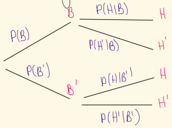
Formula

Conditional: $$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Multiplication Rule: $P(A \text{ and } B) = P(B) \cdot P(A|B)$

Multiplication Rule When Independent: $P(A \text{ and } B) = P(A) \cdot P(B)$

Note: Tree Diagram:



Notice: multiply across branches
add cases going down branches

$$P(B'|H) = \frac{P(B' \text{ and } H)}{P(H)}$$

$$= \frac{P(B') \cdot P(H|B')}{P(B') \cdot P(H|B') + P(B) \cdot P(H|B)}$$

Definition

- False Positive Rate: $P(+|D')$
- False Negative Rate: $P(-|D)$

- Sensitivity: $P(+|D)$ TRUE POSITIVE RATE
- Specificity: $P(-|D')$ TRUE NEGATIVE RATE

SET 8 - Probability Density Functions

☐ Probability

Definition

- Random Variable (r.v.): denote by X, is a function or a rule that assigns a number to each outcome of the experiment.
- Probability Density function (pdf) or Probability Distribution: is a table, formula, or graph that describes the possible values of the r.v. and the probability that each value will occur.

pdf $f(x) = P(\underset{\uparrow \text{ r.v.}}{X} = \underset{\uparrow \text{ specific value.}}{x})$

- Discrete r.v. X:
  1. $f(x) = P(X=x)$ is defined for all values of x.
  2. $f(x) = P(X=x) \geq 0$ for all values of x
  3. $\sum_x f(x) = \sum_x P(X=x) = 1$

ex.

X	0	1
$f(x)=P(X=x)$	0.75	0.25

- Cumulative Distribution Function (cdf): of a r.v. X is defined as

cdf $F(x) = P(X \leq x)$

ex. $F(c) = P(X \leq c) = \sum_{x \leq c} P(X\ x) = \sum_{x \leq c} f(x)$

ex.

X	$f(x) = P(X=x)$
0	0.75
1	0.25

pdf

X	$F(x) = P(X \leq x)$
0	0.75
1	1

cdf

Note: Last # is always 1 in cdf.

SET 9- Expectation

☐ Probability

Definition

- Expected Value of X : denoted $E(X)$, is the long-run theoretical average value of X.
- Let X be a discrete random variable w/ pdf $f(x)$. The expected value or mean value of X is given by...

$$\mu = E(X) = \sum_{\text{all } x} x \cdot f(x) = \sum_{\text{all } x} x \cdot P(X=x)$$

SET 10- Variance

☐ Probability

Rules for Expected Value

$$\left.\begin{array}{l} \circ\ \mathcal{E}(X+c) = \mathcal{E}(X) + C \\ \cdot\ \mathcal{E}(c) = C \\ \circ\ \mathcal{E}(cX) = c \cdot \mathcal{E}(X) \end{array}\right\} \mathcal{E}(aX+b) = a \cdot \mathcal{E}(X) + b$$

- Variance r.v. X w/ pdf $f(x)$:

$$V(X) = \sigma_X^2 = \sigma^2 = \mathcal{E}((X-\mu)^2)$$
$$= \sum_{\text{all } x} (x-\mu)^2 \cdot f(x)$$
$$= \sum_{\text{all } x} (x-\mu)^2 \cdot P(X=x)$$

$$V(X) = \mathcal{E}(x^2) - (\mathcal{E}(X))^2$$
$$= \mathcal{E}(X^2) - \mu^2$$

$$= \left(\sum_{\text{all } x} x^2 \cdot f(x)\right) - \mu^2$$

- Standard deviation r.v. X :

$$\sigma_X = \sigma = \sqrt{\sigma_X^2} = \sqrt{V(X)}$$

Rules for Variance and Standard Deviation

$$\left.\begin{array}{l} \circ\ V(X+c) = V(X) \\ \circ\ V(c) = 0 \\ \circ\ V(cX) = c^2 \cdot V(X) \end{array}\right\} V(aX+b) = V(aX) = a^2 \cdot V(X)$$

$$\left.\begin{array}{l} \circ\ \sigma_{X+c} = \sigma_X \\ \circ\ \sigma_c = 0 \\ \circ\ \sigma_{cX} = |c| \cdot \sigma_X \end{array}\right\} \sigma_{aX+b} = \sigma_{aX} = |a| \cdot \sigma_X$$

Rule: for r.v. X and Y → $\mathcal{E}(X+Y) = \mathcal{E}(X) + \mathcal{E}(Y)$

Rule: for r.v. X and Y (Independent) → $V(X+Y) = V(X) + V(Y)$

- $\mathcal{E}(\bar{X}) = \mu$
- $\mathcal{E}(\bar{X}) = \mathcal{E}(X)$

$$\left.\begin{array}{l} \circ\ V(\bar{X}) = V(X)/n = \sigma^2/n \\ \\ \circ\ \sigma_{\bar{X}} = \sigma_X/\sqrt{n} \end{array}\right\} \begin{array}{l} \circ\ X\text{'s are independant} \\ \circ\ \text{all have variance } \sigma_X^2 \end{array}$$

SET 11 - The Binomial Distribution

☐ Probability

Binomial Experiment :

1) Fixed number of trials, n.
2) Trial results : a success (s) or a failure (f).
3) Trials are independent and th probability of success in each trial is th same.
   $p = P(\text{success in a single trial})$
4) The random variable X counts th number of successes in n trials.

- $\binom{n}{x} = n!/x!(n-x)! = {}_nC_x = C(n,x)$

pdf for a Binomial r.v. X :

$$f(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Rules for th Binomial Distribution

- $E(X) = n \cdot p$
- $V(X) = n \cdot p(1-p)$
- $\sigma_X = \sqrt{n \cdot p(1-p)}$

Note : $p(X=c)$ use pdf formula!
$p(X \leq c)$ use cdf table.

SET 12 - The Poisson Distribution

☐ Probability

Definition

- Poisson r.v. X counts th number of events in an interval of time / length / space / etc..

$$f(x) = P(X=x) = \frac{(\lambda s)^x}{x!} e^{-\lambda s}$$

s = scaling factor : th number of basic unit measurements in our interval of interest.
λ = lamda : th average number of event occurrences in our base unit of measurement.

ex. $p(x \geq c) = 1 - p(x \leq c-1)$ ◦ Use cdf table!

Rules for th Poisson Distribution

- $E(X) = \lambda s$
- $V(X) = \lambda s$
- $\sigma_X = \sqrt{\lambda s}$

Ex. Binomial + Poisson Question

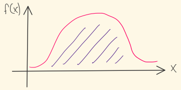
- Want to answer binomial question

n = in setup
p = solve using poisson } p is P(success in 1 trial)

→ Use poisson to find p.
→ use p in th binomial calculation

SET 13- Continuous Densities

☐ Probability



Continuous

/// area $\int_a^b f(x)\,dx$

Rules for th pdf of a continuous r.v. X

- $f(x) \geq 0$ for all x values
- The area bounded by th graph of f(x) and th x-axis is 1.
- $P(a \leq X \leq b)$ = area between $x=a$, $x=b$, $f(x)$, and th x-axis.

- If X is a continuous random variable then for a constant c
  $P(X=c)=0$.

  $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$

- $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$
  → is in th middle.

- $P(X \geq a) = 1 - P(X \leq a)$

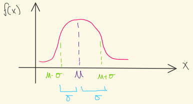
SET 14 - The Normal Distribution

□ Probability

The pdf for th normal distribution is ...

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

$-\infty < X < \infty$
μ = mean
σ = standard deviation



Normal Distribution: th bell curve, symmetric about μ, inflection points at one s.d. away from th mean (at $x = \mu \pm \sigma$).

Standardize p.5 cumulative distribution: standard normal

$\mu = 0$, $\sigma = 1$ => standard normal random variable Z

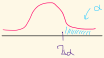
Note: X for original distribution
Z for standardized distribution

$P(Z \leq z)$ on table

$$Z = \frac{X-\mu}{\sigma}$$

shifts th mean to 0.
scales th st.dev. to 1

- z_α is th z value so that th area to th right is α.
  $P(Z \geq z_\alpha) = \alpha$ (critical value)



Notation ~ = is distributed.

- $X \sim N(\mu, \sigma)$ X is normally distributed w/ mean μ and st.dev. σ.
- $X \sim N(\mu, \sigma^2)$ X is normally distributed w/ mean μ and variance σ^2.
- $X \sim$ Binomial (n, p) X is a binomial r.v. w/ parameters n and p.
- $X \sim$ Poisson (λs) X is Poisson r.v. w/ parameters λ and s.

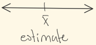
The Normal Probability Rule / Th Empiricle Rule / Th 68-95-99 Rule

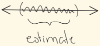
- If X is a random variable from an approximately bell-shaped distribution then...
  - $P(\mu-\sigma \le X \le \mu+\sigma) \approx 0.68$
  - $P(\mu-2\sigma \le X \le \mu+2\sigma) \approx 0.95$
  - $P(\mu-3\sigma \le X \le \mu+3\sigma) \approx 0.99$

SET 15 - Sampling

☐ Probability

A point estimate is a single valued statistic used to estimate a population parameter.





- $\bar{x}$ gives a point estimate of μ.
- s gives a point estimate of σ.

- μ = population / actual / true / theoretical value of th mean
- σ = population / actual / true / theoretical value of th standard deviation

- Random Sample:
  - All of $X_1, X_2, \dots, X_n$ have th same mean and variance.
  - Th distribution of $X_1, X_2, \dots, X_n$ all have th same shape.
  - All th random variables $X_1, X_2, \dots, X_n$ are independent of each other.

- $\bar{X} = \dfrac{X_1 + X_2}{2}$ => pdf for $\bar{X}$

SET 16 - Expectation and Variance

☐ Probability

ex. $E(2X+Y) = E(2X) + E(Y) = 2E(2X) + E(Y)$ ☆ random variables

ex. $V(2X+Y) = V(2X) + V(Y) = 2^2 V(X) + V(Y) = 4V(X) + V(Y)$ ☆ Independent random variables
→ Always add "error measurement"

Standardizing

$$\frac{\text{r.v.} - \text{expected value of r.v.}}{\text{st. dev. of r.v.}}$$

☆ $\dfrac{X-\mu}{\sigma}$ ☆ $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

SET 17 - Inferences About th Population Mean

☐ Inferential

Rule #1:
- random indepedent sample of size n
- normal distribution (μ, σ)

=> $\bar{x}$ is normal w/ μ and s.d. $\sigma/\sqrt{n}$

$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal. $\bar{X} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}$

Central Limit Theorem (CLT):

- random sample of size n
- any distribution (μ, σ)
- $n > 30$

=> $\bar{x}$ is approximately normal w/ μ and s.d. $\sigma/\sqrt{n}$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is approximately standard normal. $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$

Rule #2:

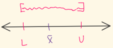
- random sample of size n
- any distribution (μ) and unknown σ.
- $n > 40$

=> $\bar{x}$ is approximately normal w/ μ and e.s.d. $s/\sqrt{n}$

$\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is approximately standard normal. $\bar{X} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Confidence Intervals

Interval Estimate



A Confidence Interval (CI):

- interval [L, U]

which gives an estimate for the population mean μ with some degree of certainty.

ex. A 95% confidence interval for μ has

$P(L \leq \mu \leq U) = 0.95$

$$[L, U] = [\underbrace{\bar{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n}}_{L}, \underbrace{\bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}}_{U}]$$

$$\boxed{(\text{estimate}) \pm (\text{c.v.})(\text{e.s.e})}$$

Note:

- Increasing the confidence level (%) => wider CI.
- Increasing n => CI narrower.
- Increasing σ => CI wider.

SE7 18 - The T Distribution

□ Probability

Rule #3:

- a random sample size n (small n)
- a normal distribution (μ) and unknown σ

$\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is T distributed w/ n-1 degrees of freedom.

Note: • normal distribution $n > 40$
• T distribution $n \leq 40$

Properties of T Random Variables

- Infinitely many T r.v. - parameter r (gamma) } T_r
- r - degrees of freedom (d.f.): positive integer
- T r.v. is continuous
- Graph of the pdf is symmetric, bell-shaped curve centered at $\mu = 0$.

- r dictates the shape of the T pdf.
  - → r increases, t_r decreases
- $T_\infty = Z$
- $t_{r,\alpha}$ $\quad P(T_r \geq t_{r,\alpha}) = \alpha$

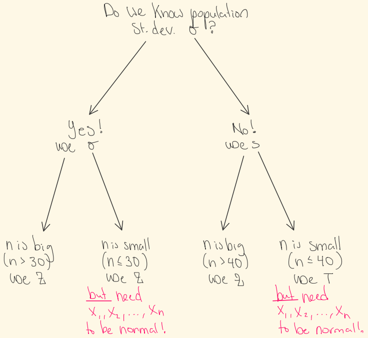
Confidence Intervals

$$[L, U] = [\bar{x} - t_{r,\alpha/2} \cdot S/\sqrt{n}, \ \bar{x} + t_{r,\alpha/2} \cdot S/\sqrt{n}]$$

$$= \bar{x} \pm t_{r,\alpha/2} \cdot S/\sqrt{n}$$

where $r = n-1$

When to use Z versus T (for $\bar{x}$)


Do we know population St. dev. σ?
Yes! use σ
No! use S
n is big (n > 30) use Z
n is small (n ≤ 30) use Z
but need $x_1, x_2, \dots, x_n$ to be normal!
n is big (n > 40) use Z
n is small (n ≤ 40) use T
but need $x_1, x_2, \dots, x_n$ to be normal!

Note: this applies to Confidence Intervals and to calculating probabilities.

SET 19 - Hypothesis Testing

- Inferential

- Null Hypothesis H_0 $\quad \mu = \mu_0$
- Alternative Hypothesis H_1 $\quad \mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$

We assume that the null hypothesis holds unless we have enough evidence to suggest the alternative hypothesis.

Two Outcomes:

- There is enough evidence to reject H_0, so we accept H_1.
- There is not enough evidence to reject H_0, so we keep believing H_0.

- Type I Error: Conclude that we reject H_0, but we shouldn't have. This happens w/ probability α. ← very serious error!
- Type II Error: Conclude there is not enough evidence to reject H_0, but we should have rejected it. This happens w/ probability β.

- We want α (the significance level) to be small.

Rule: As α (probability of Type I error) gets smaller, β (probability of Type II error) gets larger, and vice versa.

* We can reduce the probability of both types of errors by increasing the sample size n.

★ Any time H_0 is rejected we could make a Type I error. Any time H_0 is not rejected we could make a Type II error.

One tail test : $H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$ or $\mu < \mu_0$

Two tailed test : $H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

How Hypothesis Testing Works:

- Assume H_0 is true ($\mu = \mu_0$)
- Measure $\bar{x}$. Standardizing $\bar{x}$ gives the test statistic.
- Find the probability of observing a mean value as extreme as the observed $\bar{x}$ while assuming the actual mean is $\mu = \mu_0$. The probability is called the p-value.
- If this probability is small, either we observed something extremely rare, or else our assumption that $\mu = \mu_0$ was wrong. It is more likely that $\mu = \mu_0$ was wrong, so we reject H_0 assumption. If the probability is not small, our $\mu = \mu_0$ assumption was probably OK and we keep believing H_0.

The Steps for Hypothesis Testing:

1. Define the parameter to be tested.
2. Specify H_0 and H_1.
3. Specify the test statistic and identify its approximate distribution if H_0 is true.
4. Compute the observed value of the test statistic.
5. Compute the p-value.
6. Report the strength of evidence against H_0 in favour of H_1.
   - p-value small ⇒ reject H_0.
   - p-value big ⇒ Keep H_0.
7. Test H_0 at the significance level α.
   - $p\text{-value} \leq \alpha$ ⇒ reject H_0.
   - $p\text{-value} > \alpha$ ⇒ Keep H_0.

Strength of Evidence

- Very Strong Evidence if $p\text{-value} \leq 0.01$. ← reject H_0
- Strong Evidence if $0.01 < p\text{-value} \leq 0.05$.
- Moderate Evidence if $0.05 < p\text{-value} \leq 0.10$.
- Little or No Evidence if $0.10 < p\text{-value}$. ← Keep H_0

SET 20 - Hypothesis Testing of A Population Mean

☐ Inferential

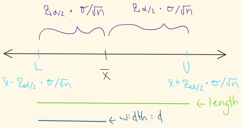
1. Testing Parameter : Testing μ = true average of < ... >.
4. Z_{obs} or T_{obs}
5. p-value $= P(Z < z)$ or $P(Z > z)$ } one-tail test
   → same inequality as H_1.
   $= P(Z \leq z) + P(Z \geq z)$ } two-tailed test
   $= 2 \cdot P(Z \leq z)$

SET 21 - Confidence Intervals and Power

☐ Inferential

$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and $\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Interval:


$z_{\alpha/2} \cdot \sigma/\sqrt{n}$
$z_{\alpha/2} \cdot \sigma/\sqrt{n}$
L
$\bar{x}$
U
$\bar{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n}$
$\bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}$
← length
← width = d

- length of interval $= 2(z_{\alpha/2} \cdot \sigma/\sqrt{n})$
- width of interval $= z_{\alpha/2} \cdot \sigma/\sqrt{n} = d$ ← margin of error.

Determine n for Sample Size

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{d}\right)^2$$

n has to be an interger.
∴ Round up!

$$n = \left(\frac{z_{\alpha/2} \cdot s}{d}\right)^2$$

SET 22 - Confidence Intervals for Population Proportions

☐ Inferential

$\hat{p} = x/n$ = # in sample w/ trait / # in sample space

where $\hat{p}$ is a point estimate for the population proportion p.

Rule #1:

- $n > 30$
- $np \geq 10$
- $n(1-p) \geq 10$

Sample proportion $\hat{p} = x/n$ is normally distributed.

- expected value of $\hat{p}$ is p
- variance $p(1-p)/n$

Rule #2:

- $n > 30$
- $np \geq 10$
- $n(1-p) \geq 10$

r.v. $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ is ~standard normal.

Confidence Interval

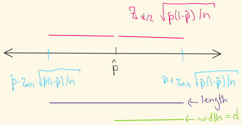
$$[L, U] = [\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

SET 23 - Sample Size and Hypothesis Tests for a Population Proportions

☐ Inferential

Interval:


$z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
$\hat{p}$
$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
$\hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
← length
← width = d

- length of interval $= 2(z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
- width of interval $= z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = d$

Determine n for Sample Size

$$n = \left(\frac{z_{\alpha/2}}{d}\right)^2 (\hat{p}(1-\hat{p}))$$

n has to be an integer
∴ Round Up!

What if We Don't Have A Good Measured Guess for $\hat{p}$?

$\hat{p} = 0.5$

$$n = \frac{(z_{\alpha/2})^2}{4d^2}$$

SET 24 - Testing Two Proportions

☐ Inferential

Want to estimate $p_1 - p_2 \Rightarrow \hat{p}_1 - \hat{p}_2 = x_1/n_1 - x_2/n_2$

Rule: n is large

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad \Big\}\ \text{standard Error!}$$

is ~ standard normal

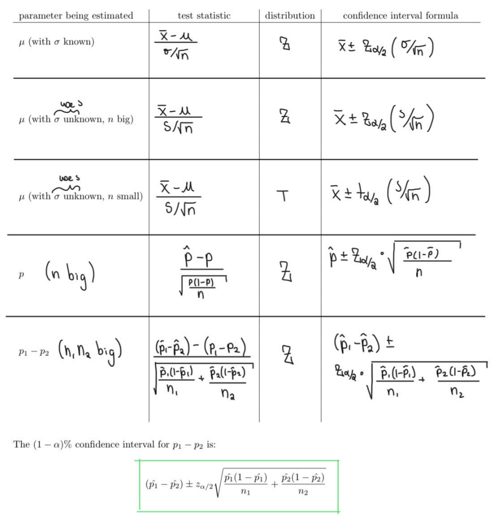
- p_1 is normally distributed
- p_2 is normally distributed

Standardize

$$Z = \frac{(\hat{p}_1 - p_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad \Big\}\ \text{estimated standard error!}$$

Review

- To standardize: $\dfrac{\text{estimate} - \text{expected value}}{\text{standard error}}$
- Confidence Interval: estimate ± (critical value)(standard error)



parameter being estimated	test statistic	distribution	confidence interval formula
μ (with σ known)	$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$	Z	$\bar{x} \pm Z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$
μ (with $\overset{\text{use } s}{\sigma}$ unknown, n big)	$\frac{\bar{x}-\mu}{s/\sqrt{n}}$	Z	$\bar{x} \pm Z_{\alpha/2}\left(s/\sqrt{n}\right)$
μ (with $\overset{\text{use } s}{\sigma}$ unknown, n small)	$\frac{\bar{x}-\mu}{s/\sqrt{n}}$	T	$\bar{x} \pm t_{\alpha/2}\left(s/\sqrt{n}\right)$
p (n big)	$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	Z	$\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$p_1 - p_2$ (n_1, n_2 big)	$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	Z	$(\hat{p}_1-\hat{p}_2) \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

The $(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is:

$$\boxed{(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

SET 25 & 26 - Comparing Two Population Means

☐ Inferential

Cases:

Z 1. We know σ_1 and σ_2, and both populations follow th normal distribution.
Z 2. Both sample sizes are large ($n_1 > 40$ and $n_2 > 40$)
T 3. Small n, equal variances: larger st.dev. / smaller st.dev. ≤ 1.4
T 4. Small n, unequal variances: larger st.dev. / smalle st.dev. > 1.4

CASE 1

- Test Statistic: $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)}}$

- Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

CASE 2

- Test Statistic: $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)}}$

σ^2 can be substituted for s^2!

- Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

σ^2 can be substituted for s^2!

CASE 3 (a normal (or near-normal) distribution)

Use the Pooled t-test

- Test Statistic: $T = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

- Degrees of freedom: $r = n_1 + n_2 - 2$

- Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

CASE 4 (a normal (or near-normal) distribution)

Use the Unpooled t-test

- Test Statistic: $T = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}}$

- Degrees of freedom:

$$r = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- Needs to be an integer
→ ROUND DOWN!

- Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

SET 27 - Paired Tests

☐ Inferential

- Pair up these numbers in a meaningful way

	1	2	3
Group 1	4	4	4
Group 2	2	4	3

Group 1 - Group 2 : 2, 0, 1 } use these numbers to find $\bar{x}_d$, s_d, and n_d

- Test Parameter: $\mu_d = \mu_1 - \mu_2$

- Test Statistic: $t = \frac{\bar{x}_d - \mu_d}{s_d / \sqrt{n}}$ $r = n-1$

- Confidence Interval: $\bar{x}_d \pm t_{r,\alpha/2} \cdot s_d / \sqrt{n}$ $r = n-1$

Note: differences need to be normally distributed!

SPECIAL NOTE: The CI doesn't have to have the same conclusion as the Hypothesis Test!

End of STAT 255 - Spring 2021