STAT 350 Help Session for Midterm 2

Chapter 7: Sampling Distributions

Parameter, statistic, and sampling distribution

- Parameter: a fixed number that describes some attribute of the population e.g., μ
- Statistic: a numerical measurement summarizes the data e.g., $\bar{x}$
- Sampling distribution: the probability distribution of a statistic e.g., distribution of $\bar{X}$

Sampling distribution of Sample Mean $\bar{X}$

Assume $X_1, X_2, \cdots, X_n$ are indep. r.v.'s with identical dist. $f_X(x)$ with finite mean μ_X and finite variance σ_X^2

- Expectation: $\mu_{\bar{X}} = E[\bar{X}] = \mu_X$, Variance: $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$, Standard Deviation: $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$

Exact Distribution of Sample Mean using Normal distribution properties

$X_1, X_2, \cdots, X_n$ are i.i.d. samples from a normal distribution with finite mean μ_X and finite variance σ_X^2, then

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\right) \text{ for any sample size } n.$$

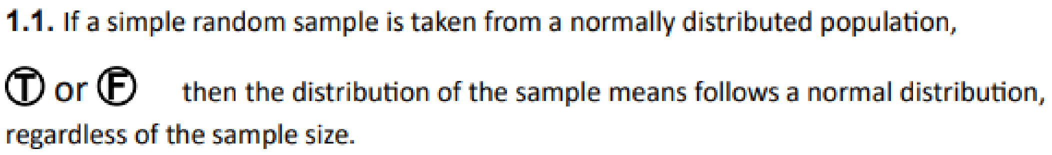
Approximated Distribution of Sample Mean using Central Limit Theorem

$X_1, X_2, \cdots, X_n$ are i.i.d. samples from some unknown population with finite mean μ_X and finite variance σ_X^2, then

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\right) \text{ for sufficiently large sample size } n.$$

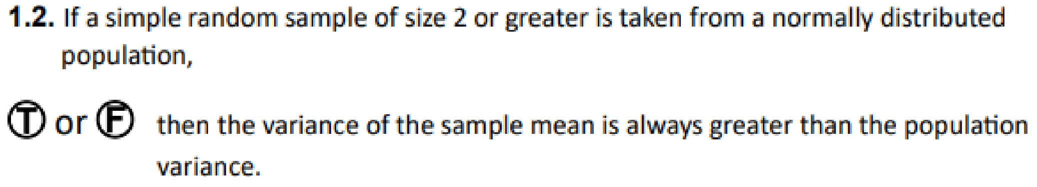
$n > 30$ is sufficient in many cases; a larger sample is required if the underlying population is far from Normal.

SPRING 2024



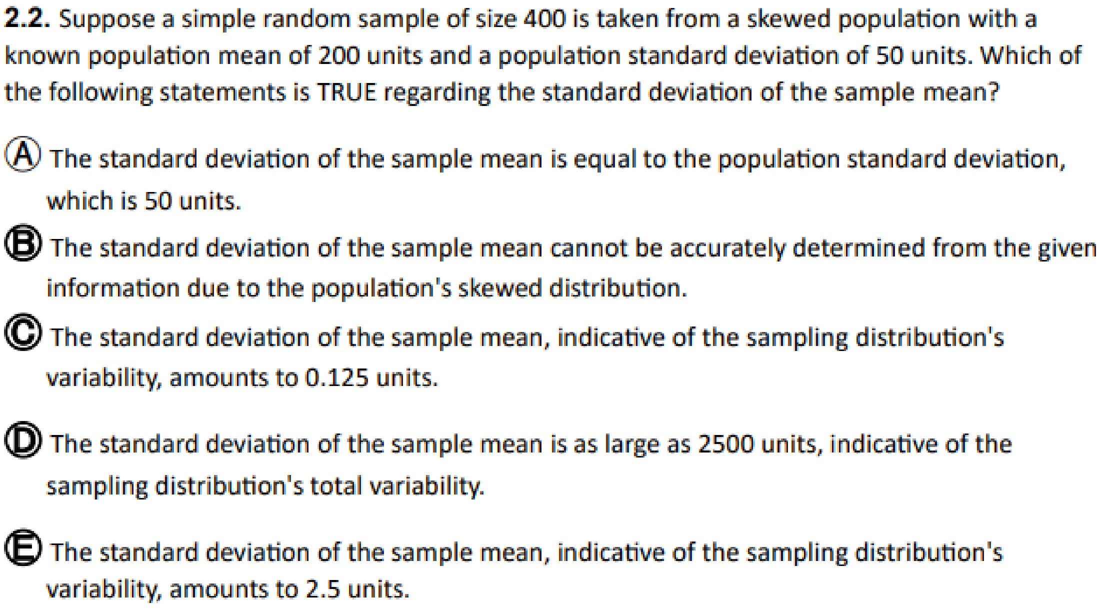
1.1. If a simple random sample is taken from a normally distributed population,

Ⓣ or Ⓕ then the distribution of the sample means follows a normal distribution, regardless of the sample size.



1.2. If a simple random sample of size 2 or greater is taken from a normally distributed population,

Ⓣ or Ⓕ then the variance of the sample mean is always greater than the population variance.



2.2. Suppose a simple random sample of size 400 is taken from a skewed population with a known population mean of 200 units and a population standard deviation of 50 units. Which of the following statements is TRUE regarding the standard deviation of the sample mean?

Ⓐ The standard deviation of the sample mean is equal to the population standard deviation, which is 50 units.

Ⓑ The standard deviation of the sample mean cannot be accurately determined from the given information due to the population's skewed distribution.

Ⓒ The standard deviation of the sample mean, indicative of the sampling distribution's variability, amounts to 0.125 units.

Ⓓ The standard deviation of the sample mean is as large as 2500 units, indicative of the sampling distribution's total variability.

Ⓔ The standard deviation of the sample mean, indicative of the sampling distribution's variability, amounts to 2.5 units.

FALL 2023 (Midterm 1)

2.2. Suppose $X_1, X_2, \dots, X_n$ is a random independent sample coming from the same (identically distributed) but unknown distribution with finite non-zero variance σ^2. **Identify the incorrect statement**.

Ⓐ For any **n**, $E[\bar{X}] = E[X_1]$.

Ⓑ $\mathbf{Var}(X_1 - X_2) \neq 0$

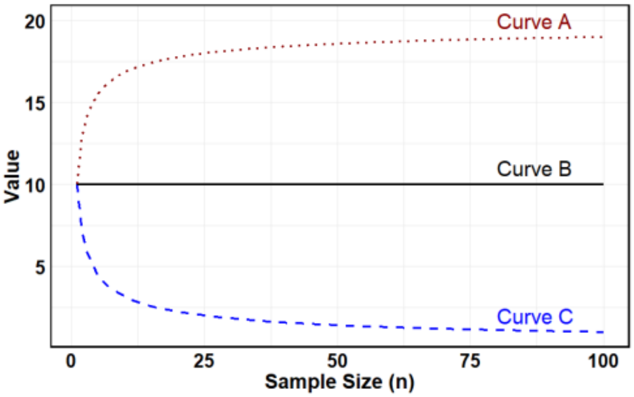
Ⓒ For any **n**, $\bar{\mathbf{X}}$ will be approximately normally distributed.

Ⓓ As **n** increases the random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ becomes approximately normally distributed with mean 0 and standard deviation 1.

Ⓔ For any **n**, $\mathbf{Var}(\bar{X}) \leq \mathbf{Var}(X_1)$.

FALL 2024

2.1. Assume $W_1, W_2, \cdots, W_n$ are **independent** samples drawn from some unknown distribution $f_W(w)$ with a **population mean** $\mu = 10$ and **population standard deviation** $\sigma = 10$. Which of the following statements is **FALSE** regarding the **distribution** of $\overline{W}$?



Ⓐ If the distribution $f_W(w)$ is heavily skewed, a larger sample is required to apply the central limit theorem.

Ⓑ **Curve A** represents the value of $sd(\overline{W})$ when the central limit theorem is not applicable.

Ⓒ **Curve B** represents the value of the $E[\overline{W}]$ for different sample sizes n.

Ⓓ **Curve C** indicates that the inference on μ_W is more accurate as the sample size increases.

SPRING 2025

1.3. Denote $\tau_n = \frac{1}{\sqrt{n}} SD(X_1 + X_2 + \cdots + X_n)$, where the X_i's are independent and identically distributed with finite variance σ^2.

Ⓣ or Ⓕ Then it follows that $\tau_3 > \tau_4$.

2.4. 📦 A delivery company, CargoSwift Logistics, operates small vans that regularly transport packages between warehouses. Each trip includes a fixed set of loading equipment (metal securing racks, crates, and straps) weighing exactly **30 lbs**. The remaining cargo consists of **sixteen** individual packages, each with weights **independently** and **identically distributed** as follows:

$$X_i \sim \text{Uniform}(a = 44, b = 56), \quad i = 1, 2, \ldots, 16.$$

Recall that $E[X_i] = \frac{a+b}{2}$ and $\text{Var}(X_i) = \frac{(b-a)^2}{12}$.

The total weight T for a typical truckload of **16 packages** is given by:

$$T = 30 + \sum_{i=1}^{16} X_i.$$

CargoSwift's delivery vans have a maximum safe weight load capacity of **850 lbs**. Select the correct code to calculate the approximate probability that a randomly selected van containing **16 packages** would **exceed** the safe weight load.

Ⓐ `pt(1.4434, df = 15, lower.tail = FALSE)`

Ⓑ `pnorm(1.4434, lower.tail = FALSE)`

Ⓒ `pt(3.6084, df = 15, lower.tail = FALSE)`

Ⓓ `pnorm(3.6084, lower.tail = FALSE)`

Ⓔ `pt(5.7735, df = 15, lower.tail = FALSE)`

Ⓕ `pnorm(5.7735, lower.tail = FALSE)`

Ⓖ `punif(850, min = 704, max = 896, lower.tail = FALSE)`

Chapter 8: Experimental Design

Sources of Data

- Observational study: observe and record the outcomes without making any active interventions
- Experimental study: manipulate one or more variables to observe their effect on one or more variables

Components of Experiments

- Experimental unit: objects being studied in an experiment (e.g., indoor cats)
- Factor: a variable manipulated to see if and how it influences the response variable (e.g., litter types)
- Level: the different values that a factor can take (e.g., three distinct litter types: bentonite, casava, wood pallets)
- Treatment: manipulation of factors (e.g., indoor cats use a specific type of litter)
- Response variable: the outcome of an experiment (e.g., time interval between litter box visits)

Principles of Experimental Design

- Control: keep other conditions the same except for the variable of interest. Use blocking if extraneous variables cannot be directly controlled.
- Randomize: experimental units are randomly assigned to the treatments to avoid bias.
- Replication: use enough experimental units or repeat experiments per treatment to reduce variation (=errors)

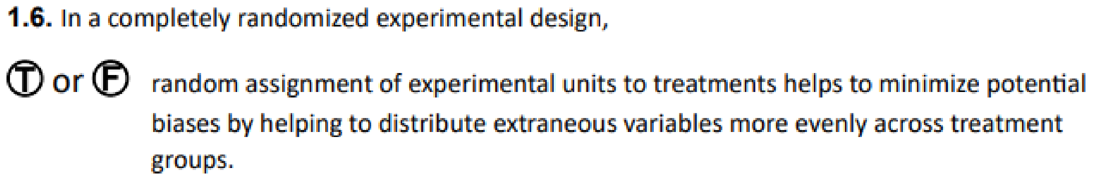
Experimental Design Graphs: refer to the graphs in lecture slides

- Completely Randomized Design: all experimental units are randomly allocated to different treatments.
- Randomized Block Design: blocks are formed based on similar characteristics before random assignment.
- Matched Pairs Design: each experimental unit is matched/paired with another similar experimental unit.

Sampling Methods

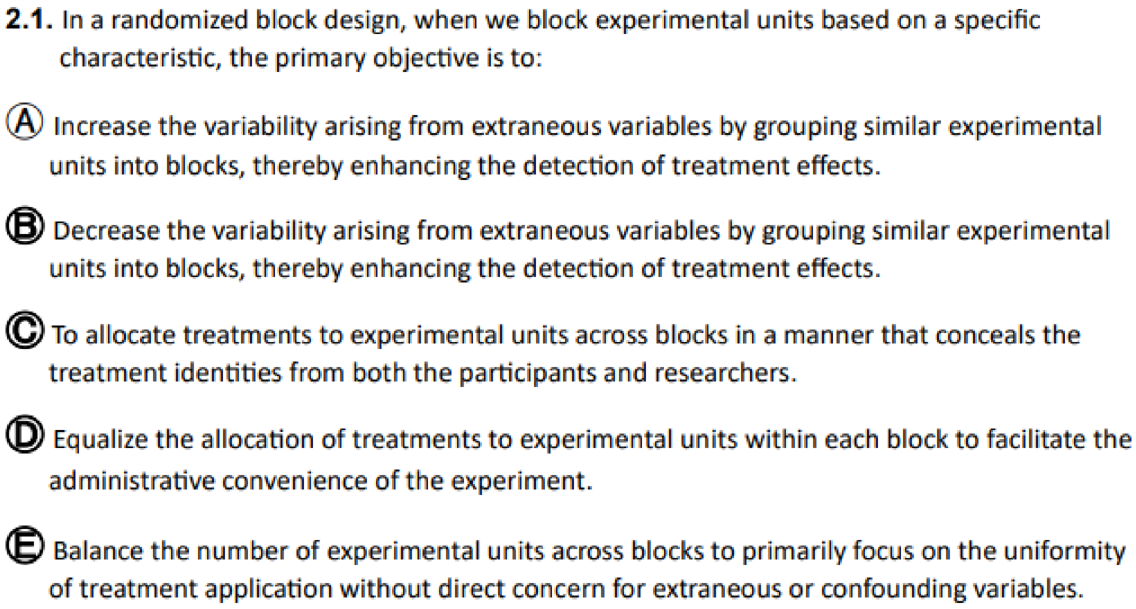
- (AVOID) convenience sample or voluntary response sample – highly possible to cause sampling bias
- (Preferred) Simple Random Sampling, Stratified Random Sampling – less likely to see sampling bias

SPRING 2024



1.6. In a completely randomized experimental design,

Ⓣ or Ⓕ random assignment of experimental units to treatments helps to minimize potential biases by helping to distribute extraneous variables more evenly across treatment groups.



2.1. In a randomized block design, when we block experimental units based on a specific characteristic, the primary objective is to:

Ⓐ Increase the variability arising from extraneous variables by grouping similar experimental units into blocks, thereby enhancing the detection of treatment effects.

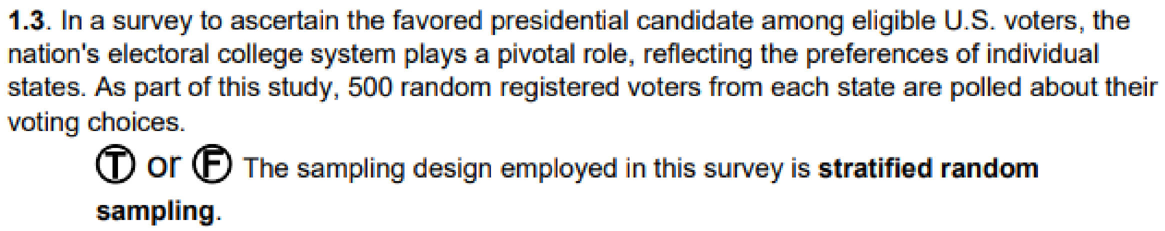
Ⓑ Decrease the variability arising from extraneous variables by grouping similar experimental units into blocks, thereby enhancing the detection of treatment effects.

Ⓒ To allocate treatments to experimental units across blocks in a manner that conceals the treatment identities from both the participants and researchers.

Ⓓ Equalize the allocation of treatments to experimental units within each block to facilitate the administrative convenience of the experiment.

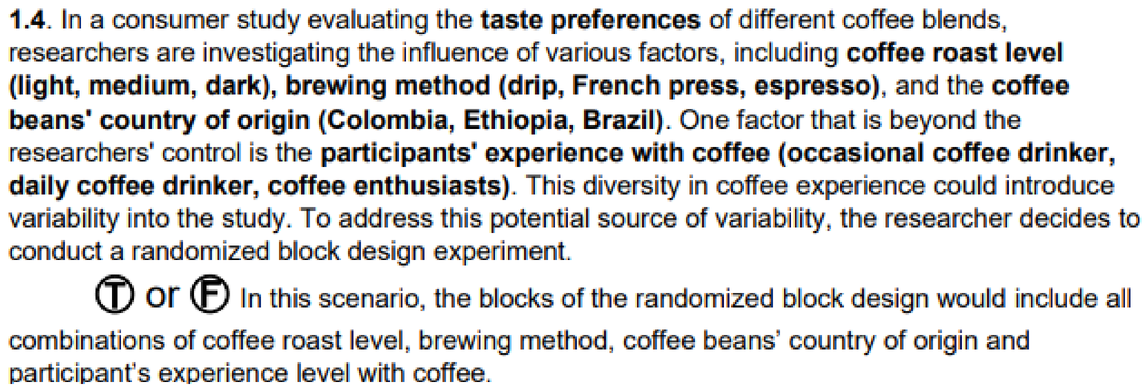
Ⓔ Balance the number of experimental units across blocks to primarily focus on the uniformity of treatment application without direct concern for extraneous or confounding variables.

FALL 2023



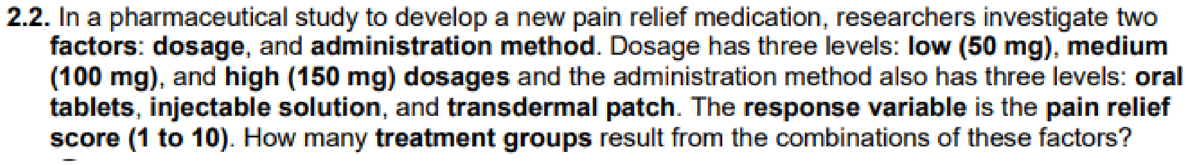
1.3. In a survey to ascertain the favored presidential candidate among eligible U.S. voters, the nation's electoral college system plays a pivotal role, reflecting the preferences of individual states. As part of this study, 500 random registered voters from each state are polled about their voting choices.

Ⓣ or Ⓕ The sampling design employed in this survey is **stratified random sampling**.



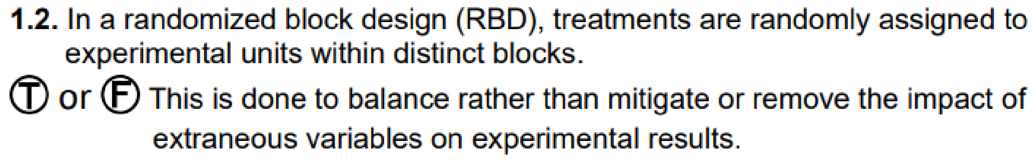
1.4. In a consumer study evaluating the **taste preferences** of different coffee blends, researchers are investigating the influence of various factors, including **coffee roast level (light, medium, dark), brewing method (drip, French press, espresso)**, and the **coffee beans' country of origin (Colombia, Ethiopia, Brazil)**. One factor that is beyond the researchers' control is the **participants' experience with coffee (occasional coffee drinker, daily coffee drinker, coffee enthusiasts)**. This diversity in coffee experience could introduce variability into the study. To address this potential source of variability, the researcher decides to conduct a randomized block design experiment.

Ⓣ or Ⓕ In this scenario, the blocks of the randomized block design would include all combinations of coffee roast level, brewing method, coffee beans' country of origin and participant's experience level with coffee.



2.2. In a pharmaceutical study to develop a new pain relief medication, researchers investigate two **factors**: **dosage**, and **administration method**. Dosage has three levels: **low (50 mg)**, **medium (100 mg)**, and **high (150 mg) dosages** and the administration method also has three levels: **oral tablets**, **injectable solution**, and **transdermal patch**. The **response variable** is the **pain relief score (1 to 10)**. How many **treatment groups** result from the combinations of these factors?

SPRING 2025

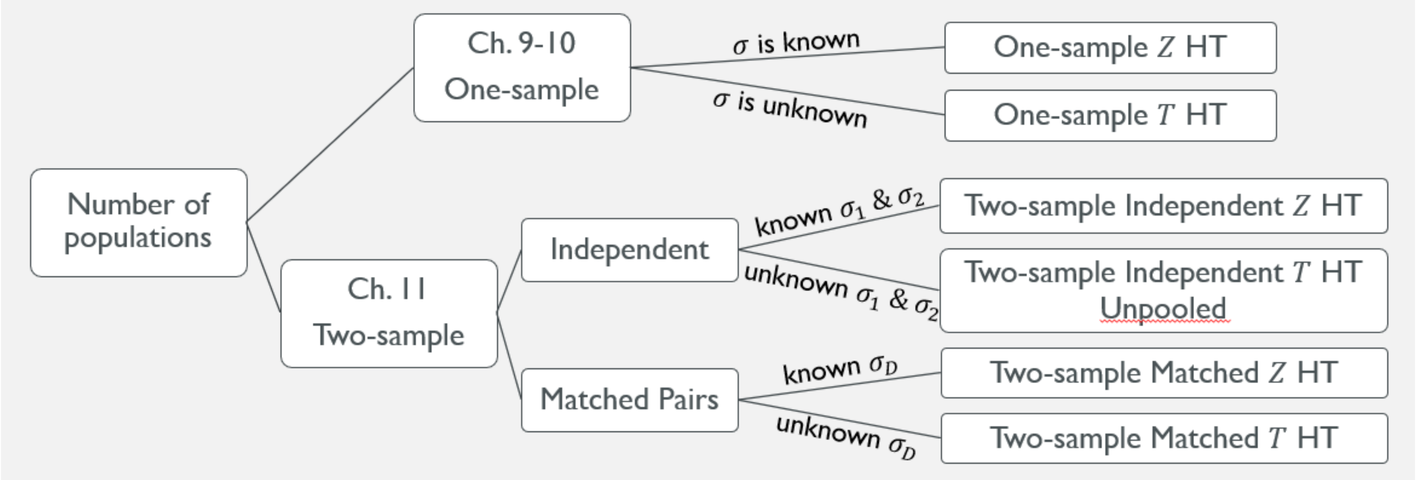


1.2. In a randomized block design (RBD), treatments are randomly assigned to experimental units within distinct blocks.

Ⓣ or Ⓕ This is done to balance rather than mitigate or remove the impact of extraneous variables on experimental results.

Identification of Correct Hypotheses

For better understanding, review how the sampling distribution of a point estimator (e.g., $\bar{d}$) is derived for each case.



Four Steps of Hypothesis Testing

STEP 1: Define parameter(s) in the context of problem.

STEP 2: State the null and alternative hypotheses – use PARAMETERS and NUMBERS only!

H_0: ________________________________ and H_a: ________________________________

STEP 3: Calculate the test statistic (z or t) and find the p-value. $z =$ _______ (or $t =$ _______). p-value = ________

STEP 4: Make the decision and conclude in the context of the problem – choose one based on your p-value and α

If p-value $\leq \alpha$: The p-value = ____ $\leq \alpha =$ ____ therefore we have evidence to reject the null hypothesis H_0. The data does give support to claim that the (*interpret H_a in the context of the problem*).

If p-value $> \alpha$: The p-value = ____ $> \alpha =$ ____ therefore we do NOT have evidence to reject the null hypothesis H_0. The data does NOT give support to claim that the (*interpret H_a in the context of the problem*).

Confidence Interval/Bound, p-values by the type of Alternative Hypothesis

Alternative Hypothesis	Confidence	p-value (z_{ts})	p-value (t_{ts})
$H_a: \mu > \mu_0$	Lower Bound	$P(Z > z_{ts})$	$P(T > t_{ts})$
$H_a: \mu < \mu_0$	Upper Bound	$P(Z < z_{ts})$	$P(T < t_{ts})$
$H_a: \mu \neq \mu_0$	Interval	$2 * P(Z > \lvert z_{ts}\rvert)$	$2 * P(T > \lvert t_{ts}\rvert)$

Four Different Hypothesis Testing Procedures

One-sample z

Parameter: μ

Hypothesized value: μ_0 – assumed mean value

Test statistic: $z_{ts} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Confidence Interval: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Confidence Lower Bound: $\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}$

One-sample t

Parameter: μ

Hypothesized value: μ_0 – assumed mean value

Test statistic: $t_{ts} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ df = n-1$

Confidence Interval: $\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$

Confidence Lower Bound: $\bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}}$

Two-sample Independent (known variances)

Parameters: μ_A and μ_B

Hypothesized value: Δ_0 – assumed diff. in pop. Means

Test statistic: $z_{ts} = \frac{\bar{x}_A-\bar{x}_B-\Delta_0}{SE},\ SE = \sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}$

Confidence Interval: $\bar{x}_A - \bar{x}_B \pm z_{\alpha/2}\ SE$

Confidence Lower Bound: $\bar{x}_A - \bar{x}_B - z_{\alpha/2} * SE$

Two-sample Independent (UNPOOLED)

Parameters: μ_A and μ_B

Hypothesized value: Δ_0 – assumed diff. in pop. Means

Test statistic: $t_{ts} = \frac{\bar{x}_A-\bar{x}_B-\Delta_0}{s_{unpool}},\ s_{unpool} = \sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}$

Welch-Satterthwaite **approx.** DF: $\nu = \frac{\left(\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}\right)^2}{\frac{1}{n_A-1}\left(\frac{s_A^2}{n_A}\right)^2+\frac{1}{n_B-1}\left(\frac{s_B^2}{n_B}\right)^2}$

Confidence Interval: $\bar{x}_A - \bar{x}_B \pm t_{\alpha/2,\nu}\ s_{unpool}$

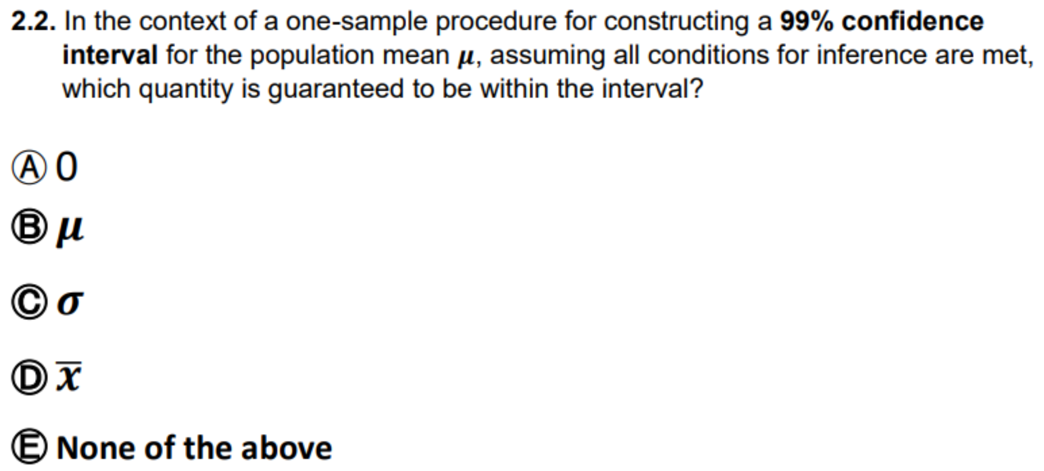
Confidence Lower Bound: $\bar{x}_A - \bar{x}_B - t_{\alpha,\nu} * s_{unpool}$

Two-sample Matched Pair (known σ_d)	**Two-sample Matched Pair (unknown σ_d)**
Parameter: μ_D – population mean of differences Hypothesized value: Δ_0 – assumed mean of diffs Test statistic: $z_{ts} = \frac{\bar{d}-\Delta_0}{\sigma_d/\sqrt{n}}$ Confidence Interval: $\bar{d} \pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}}$ Confidence Lower Bound: $\bar{d} - z_{\alpha}\frac{\sigma_d}{\sqrt{n}}$	Parameter: μ_D – population mean of differences Hypothesized value: Δ_0 – assumed mean of diffs Test statistic: $t_{ts} = \frac{\bar{d}-\Delta_0}{s_D/\sqrt{n}}, df = n-1$ Confidence Interval: $\bar{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$ Confidence Lower Bound: $\bar{d} - t_{\alpha,n-1}\frac{s_d}{\sqrt{n}}$

R Code: Normal Distribution (z)	**R Code: t Distribution**

pnorm(test.statistic, lower.tail=FALSE)
qnorm(alpha, lower.tail=FALSE)
	pt(test.statistic, df, lower.tail = FALSE)
	qt(alpha, df, lower.tail=FALSE)

FALL 2024



2.2. In the context of a one-sample procedure for constructing a **99% confidence interval** for the population mean $\boldsymbol{\mu}$, assuming all conditions for inference are met, which quantity is guaranteed to be within the interval?

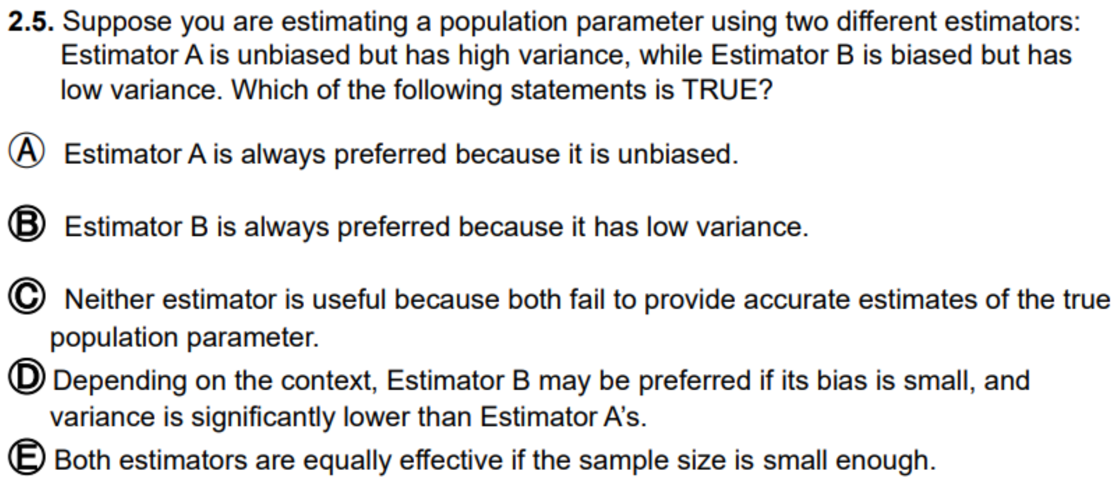
Ⓐ 0

Ⓑ $\boldsymbol{\mu}$

Ⓒ $\boldsymbol{\sigma}$

Ⓓ $\bar{\boldsymbol{x}}$

Ⓔ **None of the above**



2.5. Suppose you are estimating a population parameter using two different estimators: Estimator A is unbiased but has high variance, while Estimator B is biased but has low variance. Which of the following statements is TRUE?

Ⓐ Estimator A is always preferred because it is unbiased.

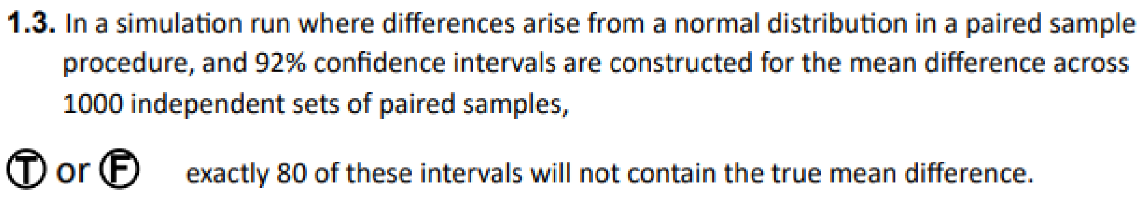
Ⓑ Estimator B is always preferred because it has low variance.

Ⓒ Neither estimator is useful because both fail to provide accurate estimates of the true population parameter.

Ⓓ Depending on the context, Estimator B may be preferred if its bias is small, and variance is significantly lower than Estimator A's.

Ⓔ Both estimators are equally effective if the sample size is small enough.

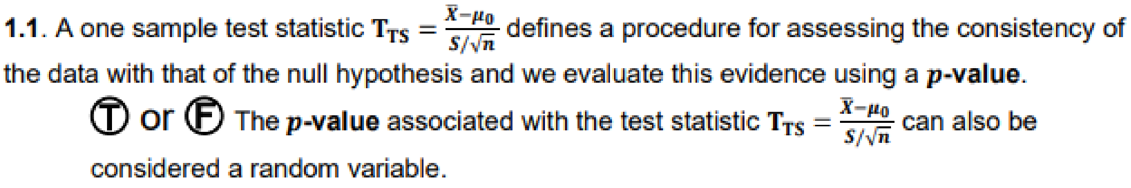
SPRING 2024



1.3. In a simulation run where differences arise from a normal distribution in a paired sample procedure, and 92% confidence intervals are constructed for the mean difference across 1000 independent sets of paired samples,

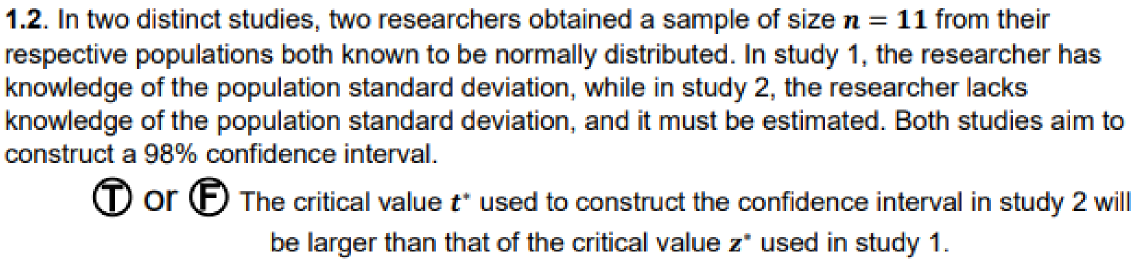
Ⓣ or Ⓕ exactly 80 of these intervals will not contain the true mean difference.

FALL 2023



1.1. A one sample test statistic $\mathbf{T_{TS}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ defines a procedure for assessing the consistency of the data with that of the null hypothesis and we evaluate this evidence using a ***p*-value**.

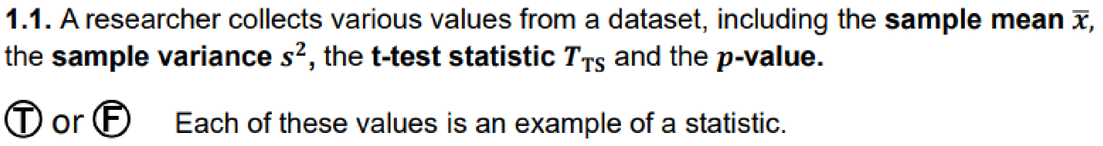
Ⓣ or Ⓕ The ***p*-value** associated with the test statistic $\mathbf{T_{TS}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ can also be considered a random variable.



1.2. In two distinct studies, two researchers obtained a sample of size $n = 11$ from their respective populations both known to be normally distributed. In study 1, the researcher has knowledge of the population standard deviation, while in study 2, the researcher lacks knowledge of the population standard deviation, and it must be estimated. Both studies aim to construct a 98% confidence interval.

Ⓣ or Ⓕ The critical value t^* used to construct the confidence interval in study 2 will be larger than that of the critical value z^* used in study 1.

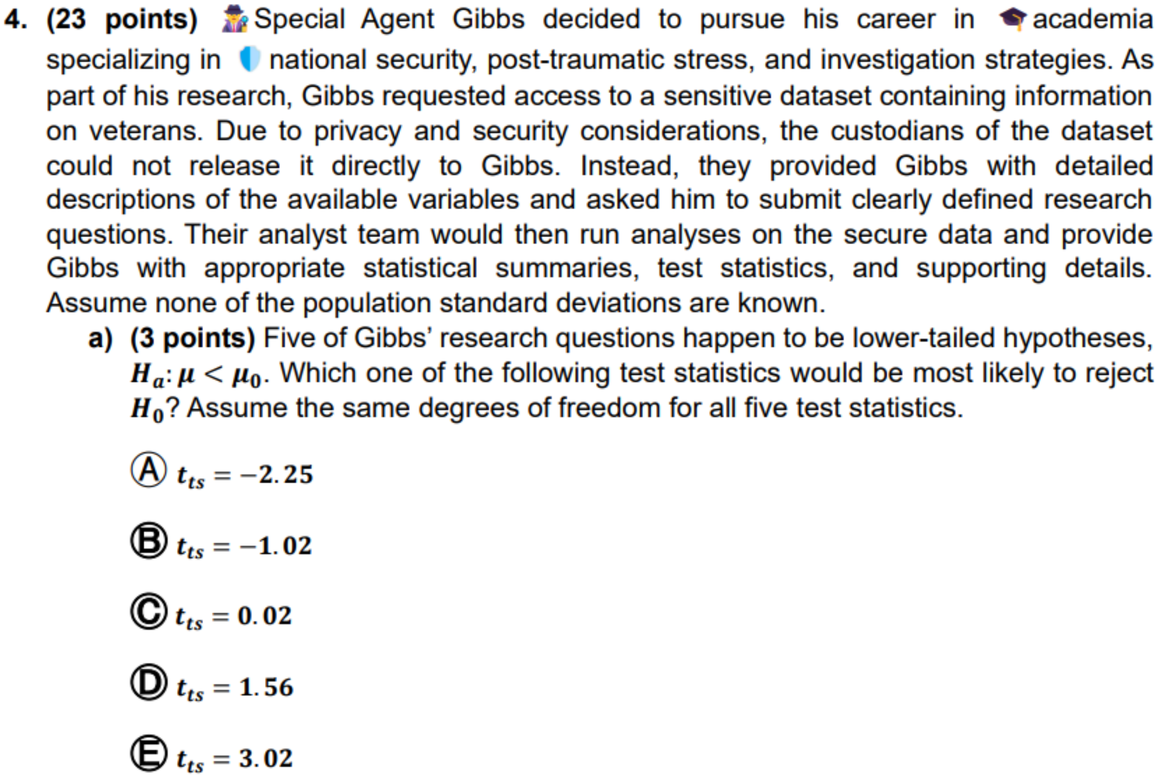
FALL 2024



1.1. A researcher collects various values from a dataset, including the **sample mean** $\bar{x}$**,** the **sample variance** s^2**,** the **t-test statistic** T_{TS} and the ***p*-value.**

Ⓣ or Ⓕ Each of these values is an example of a statistic.

SPRING 2025



4. (23 points) 🕵️Special Agent Gibbs decided to pursue his career in 🎓academia specializing in 🛡️ national security, post-traumatic stress, and investigation strategies. As part of his research, Gibbs requested access to a sensitive dataset containing information on veterans. Due to privacy and security considerations, the custodians of the dataset could not release it directly to Gibbs. Instead, they provided Gibbs with detailed descriptions of the available variables and asked him to submit clearly defined research questions. Their analyst team would then run analyses on the secure data and provide Gibbs with appropriate statistical summaries, test statistics, and supporting details. Assume none of the population standard deviations are known.

a) (3 points) Five of Gibbs' research questions happen to be lower-tailed hypotheses, $H_a: \mu < \mu_0$. Which one of the following test statistics would be most likely to reject H_0? Assume the same degrees of freedom for all five test statistics.

Ⓐ $t_{ts} = -2.25$

Ⓑ $t_{ts} = -1.02$

Ⓒ $t_{ts} = 0.02$

Ⓓ $t_{ts} = 1.56$

Ⓔ $t_{ts} = 3.02$

Additional Topics

Pooled vs Unpooled Estimator for two independent samples

Pooled	Unpooled
Assume equal variances (i.e., $\sigma_A^2 = \sigma_B^2$) $s_{pool}^2 = \left[\frac{n_A-1}{n_A+n_B-2}\right] s_A^2 + \left[\frac{n_B-1}{n_A+n_B-2}\right] s_B^2$ ☺ Simple degrees of freedom, $n_A + n_B - 2$ ☺ Test statistic is from a exact t distribution ☹ Bad when sample sizes are very different or $\sigma_A^2 \neq \sigma_B^2$	Assume unequal variances (i.e., $\sigma_A^2 \neq \sigma_B^2$) $s_{unpool}^2 = \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}$ ☺ Prevent serious errors by assuming $\sigma_A^2 = \sigma_B^2$ ☹ Complex degrees of freedom calculation (in decimals) ☹ Test statistic is not exactly t distribution

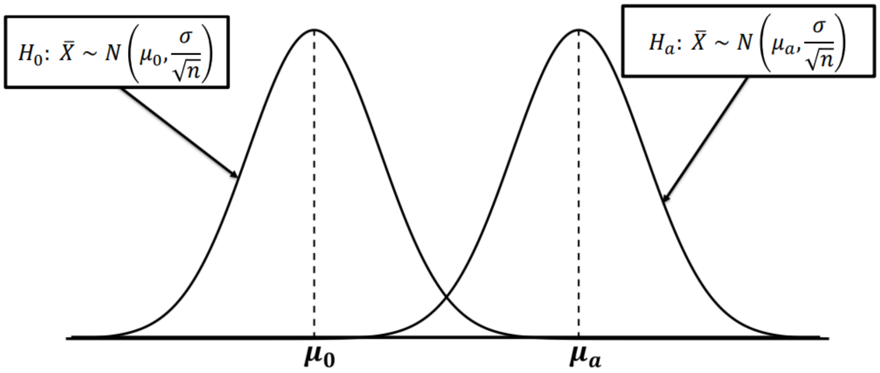
Hypothesis Testing Error and Power

- Type I Error (α): the probability of rejecting the null hypothesis H_0 incorrectly when H_0 is true (false positive)
- Type II Error (β): the probability of not rejecting the null hypothesis H_0 even though H_0 is false (false negative)
- Power $(1-\beta)$: the probability that a test correctly rejects a false null hypothesis (true positive)

Power Calculation (visuals help a lot)

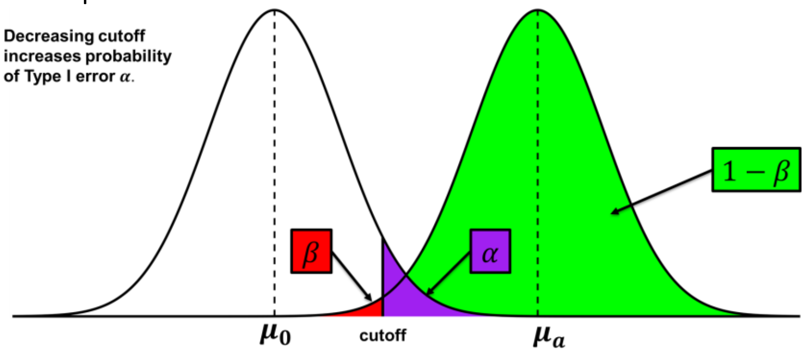
- Identify the parameter of interest, hypotheses, and detectable difference μ_a.
- Find the sampling distributions of a point estimate (e.g., $\bar{X}$ or $\bar{X}_A - \bar{X}_B$) under H_0 and H_a.
- Find a cutoff value using the sampling distribution under H_0 and a given α
- Calculate the power using the sampling distribution under H_a

Example: Null and Alternative sampling distributions when σ known under upper-tail HT,



H_0: $\bar{X} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$
H_a: $\bar{X} \sim N\left(\mu_a, \frac{\sigma}{\sqrt{n}}\right)$
μ_0
μ_a

and its testing errors and power:



Decreasing cutoff increases probability of Type I error α.
$1-\beta$
β
α
μ_0
cutoff
μ_a

SPRING 2024

1.4. When the significance level ($\boldsymbol{\alpha}$) of a statistical test is reduced while holding all other factors constant,

Ⓣ or Ⓕ the power of the test increases.

1.5. In a two-sample independent t-test using the Welch procedure to account for unequal variances between groups,

Ⓣ or Ⓕ the test statistic is assumed to adhere to an exact t-distribution, provided the assumption of normality holds.

2.3. When estimating the difference between two population means using confidence intervals, i f a researcher incorrectly uses a pooled variance estimator under the false assumption of eq ual variances, despite the populations having unequal variances, how does this affect the ma rgin of error for the confidence interval?

Ⓐ The margin of error is unaffected, as the pooled estimator adjusts for variance differences.

Ⓑ The margin of error decreases, reflecting an underestimated standard error due to the assumption violation.

Ⓒ The margin of error increases, reflecting an overestimated standard error due to the assumption violation.

Ⓓ The margin of error may inaccurately reflect the true variability, underestimating or overestimating it based on the sample sizes and actual variances.

Ⓔ The margin of error becomes zero, indicating a failure of the pooled estimator to account for variance differences.

FALL 2023

1.5. The power associated with a statistical hypothesis test is stated to be **95%**.

Ⓣ or Ⓕ This indicates that the test has a **95%** sensitivity to detect the specific effect in the study when that effect is present.

2.1. Which of the following statements is accurate regarding hypothesis tests?

Ⓐ Type I Error can be considered as the rejection of the null hypothesis when the alternative hypothesis is true.

Ⓑ Power can be considered as the probability of rejecting the null hypothesis in favor of the alternative when the alternative is false.

Ⓒ When the probability of Type II Error decreases, power increases.

Ⓓ Type II Error can be considered as the rejection of the null hypothesis when the null hypothesis is true.