Module 4

Statistical Experiments

Observations of an Early Fan of Statistics

(Mark Twain ….. who blamed this on Benjamin Disraeli):

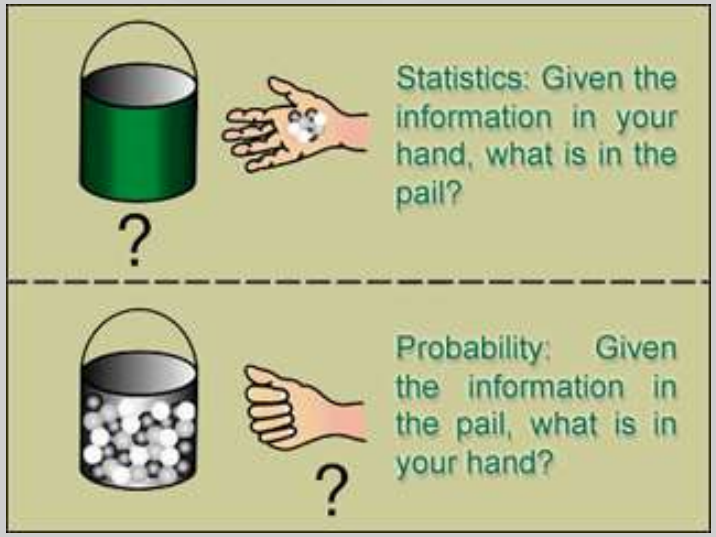
"There are ***Three Kinds of Lies – Lies, Damned Lies and Statistics!"*** ☺ ☺ ☺

Probability:

- *From population to sample (deduction)*

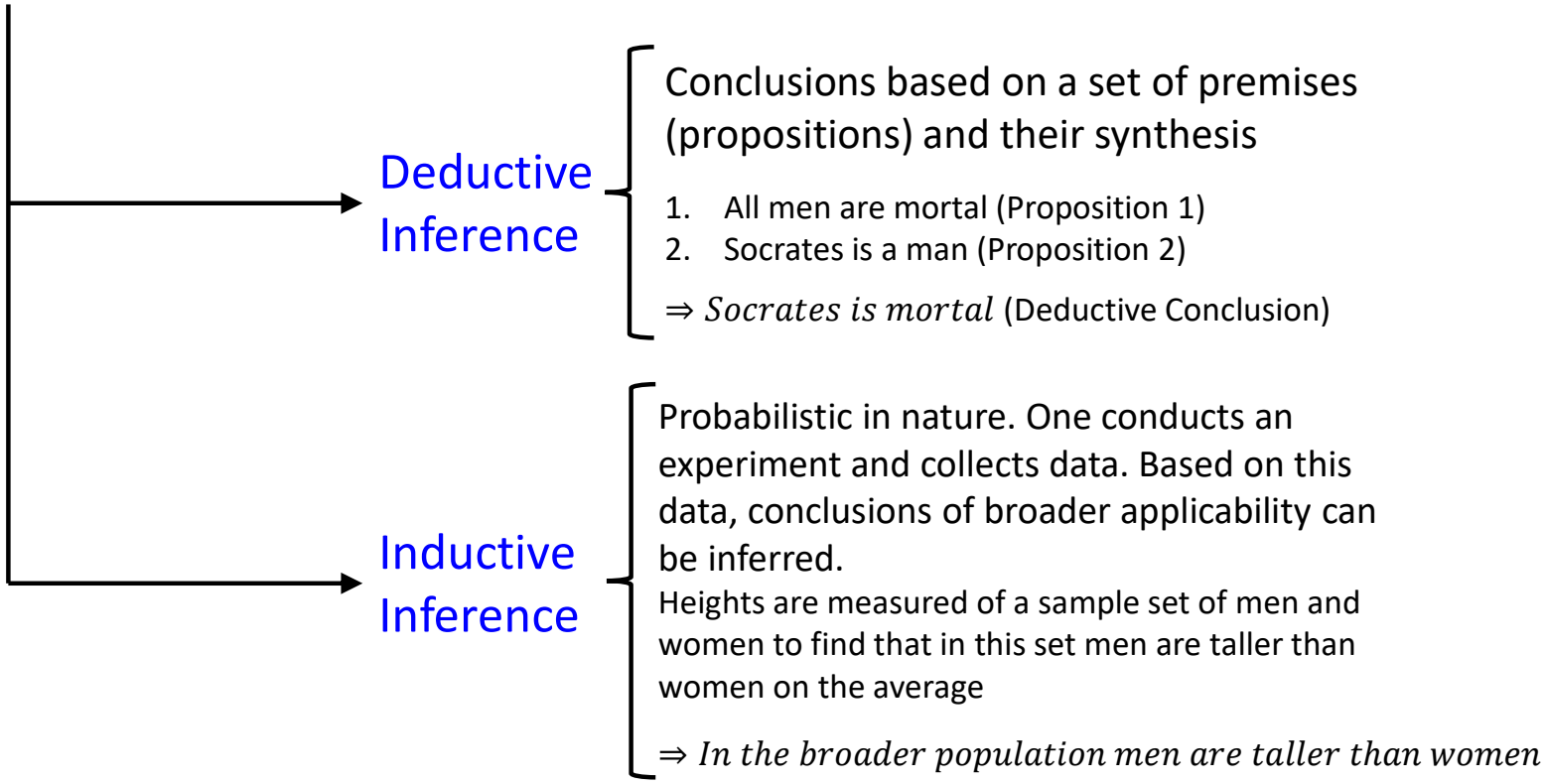
Statistics:

- *From sample to the population (induction)*



Taken from - *https://www4.stat.ncsu.edu/~reiland/courses/st511/sampling%20distributions%20and%20CLT.pptx*
Very interesting set of slides! Please do read when you have time

Inferences



Main Theme of this Module: "Draw conclusions based on inductive reasoning from data obtained from statistical experiments"

Summary of Useful Distributions for Statistics

Normal Distribution $X \sim \mathcal{N}(\mu, \sigma^2)$

PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty < x < \infty$

MGF $M(t) = E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}$ $-\infty < t < \infty$

Mean $E(X) = \mu$

Variance $Var(X) = \sigma^2$

Chi-Square Distribution $X_r \sim \chi^2(r)$ $r = 1, 2 \ldots$

r degrees of freedom; sum of the squares of r independent $\mathcal{N}(0,1)$ random variables

PDF $f_X(x) = \frac{1}{2} \frac{\left(\frac{x}{2}\right)^{\left(\frac{r}{2}\right)-1}}{\Gamma\left(\frac{r}{2}\right)} e^{-\frac{x}{2}}$ $x \geq 0$

$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$

MGF $M(t) = E(e^{tX}) = (1-2t)^{-\frac{r}{2}}$ $-\infty < t < \infty$

Mean $E(X) = \mu = r$

Variance $Var(X) = \sigma^2 = 2r$

Definition of Terms used in Statistical Sampling:

Random Sample: Consider a population $\mathfrak{P}$ from which we pick n samples randomly to get the n random variables $X_1, X_2, \ldots\ldots X_n$. These constitute a random sample if they are *independent and identically distributed* (*i.i.d*) random variables.

Note the statistical techniques can be applied to draw proper inferences only if the samples $X_1, X_2, \ldots\ldots X_n$ are drawn in a random manner from $\mathfrak{P}$

Definition of Terms used in Statistical Sampling:

Statistic: A *statistic* is a function of the observations made from a random sample.

A statistic is any quantity computed from values in a sample which is considered for a *statistical purpose*. It is a random variable because it depends on randomly selected samples.

Statistical purposes include things like estimating a population parameter, describing a sample, or evaluating a hypothesis. The average (or mean) of sample values is a statistic as well.

See example in next slide

Example: Random Sample and Statistic

Want to find out what proportion p of 10,000 employees will buy subscription to a Health Facility.

Since surveying all 10,000 may not be feasible, HR picks 100 employees randomly, irrespective of gender, age, department etc. for the survey and finds that 71 of those 100 people will buy subscriptions.

Therefore, $\hat{p} = 0.71$ is a reasonable estimate of p

Note that if we had chosen a different set of people then the value of $\hat{p}$ may have been slightly different.

The statistic $\hat{p}$ is therefore itself random as it a function of the random sample

Sample Statistics: Consider the random sample $X_1, X_2, \ldots\ldots X_n$

Some frequently used sample statistics are the following –

1. Sample Mean $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots\ldots + X_n}{n}$

2. Sample Variance $S^2 := \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$

This is referred to as an Unbiased Estimator of the Population Variance σ^2.
(See next slide for details)

3. Sample Standard Deviation $S = \sqrt{S^2}$

Variance of the sum = Sum of the variances

Bienayme's Formula: If the random variables $X_1, X_2, \ldots\ldots X_n$ are independent, then $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$

$X_1, X_2, \ldots\ldots X_n$ are i.i.d random variables, each with mean μ and variance σ^2

Why call S^2 (as defined in the last slide) an *Unbiased Estimator* of the variance?

More specifically, why is $\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$ an Unbiased Estimator of σ^2 whereas $\frac{1}{(n)}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is not (i.e., it is a biased estimator of σ^2)?

Consider $\widehat{S^2} = \frac{1}{(n)}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (Uncorrected) Sample Variance

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the Sample Mean with $\bar{X} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \Rightarrow \sum_{i=1}^{n}(X_i - \mu) = n(\bar{X} - \mu)$

$$E\left(\widehat{S^2}\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i - \mu\right) - \left(\bar{X} - \mu\right)\right)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}\left(\bar{X} - \mu\right)\sum_{i=1}^{n}\left(X_i - \mu\right) + \left(\bar{X} - \mu\right)^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2\right] - E\left[\left(\bar{X} - \mu\right)^2\right] = \sigma^2 - E\left[\left(\bar{X} - \mu\right)^2\right]$$

$\frac{n\sigma^2}{n}$

$$= \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2$$

Biased Estimator as it depend on the n, the number of samples

$\bar{X}$ (sample mean) is still a random variable

$(X_i \perp X_j \quad E(X_i - \mu) = 0$

$$= E\left[\left\{\frac{X_1 + .. + X_n}{n} - \mu\right\}^2\right] = \frac{E\left[\left\{(X_1 - \mu) + ... + (X_n - \mu)\right\}^2\right]}{n^2}$$

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

This is the Variance of the Sampled Mean
Will be useful later

Now consider $S^2 := \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$$E\left(S^2\right) = E\left[\frac{1}{(n-1)}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]$$
$$= \frac{n}{(n-1)} E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]$$
$$= \frac{n}{(n-1)}\left(1 - \frac{1}{n}\right)\sigma^2 = \sigma^2$$

From previous slide!

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2$$

The Unbiased Estimator does not show any dependence on n, the number of samples being considered. This makes it better to use than the biased estimator which depends on the number of samples and converges to the true value of the variance only when the number of samples tends to infinity.

Summarizing the important points:

$X_1, X_2, \ldots\ldots X_n$ are i.i.d random variables, each with mean μ and variance σ^2

Sample Mean $\bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots\ldots + X_n}{n}$ $\qquad E\{\bar{X}\} = \mu$

Variance of the Sample Mean $\qquad Var\{\bar{X}\} = E\{(\bar{X} - \mu)^2\} = \frac{\sigma^2}{n}$

Sample Variance (Unbiased) $\qquad S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$ $\qquad E\{S^2\} = \sigma^2$

Sample Variance (Uncorrected) $\qquad \hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ $\qquad E\{\hat{S}^2\} = \left(1 - \frac{1}{n}\right)\sigma^2$

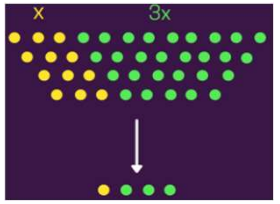
$$E\{\hat{S}^2\} = E\{S^2\} - Var\{\bar{X}\}$$

Sampling Strategies

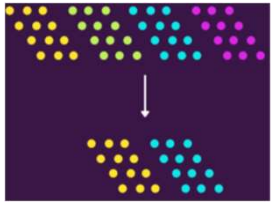
Lot of money may ride on how you sample, e.g. in "Nielsen Ratings

Random Sampling: Every member of the population equally likely to be chosen for sampling

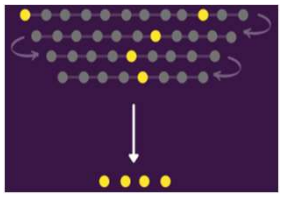
Stratified Sampling: Population classified into subpopulations based on some characteristic, sampled randomly within the sub-populations in the same proportions as in the original population.
This is used commonly in pattern classification and machine learning applications



Clustered Sampling: Population classified into subgroups such that each subgroup encompasses all the features of the entire population. The clusters are then randomly selected as a group, rather than individually, to form the sample set



Systematic Sampling: A pre-defined strategy is used to pick the members of the sample set from the whole population, e.g. every 10^{th} (randomly decided) member of the population is sampled sequentially.



Sampling Distributions:

These are the probability distributions of a statistic and are a consequence of the real outcomes of a statistical experiment.
The commonly used ones are -

- Standard normal distribution arises naturally for means or as a consequence of the Central Limit Theorem (CLT) with known mean μ and variance σ^2
- χ^2 distribution arises from the sum of the squares of normally distributed random variables which have zero mean and unit variance, i.e. $\mathcal{N}(0,1)$
- t distribution used when the population variance σ^2 is not known and only the sample variance S^2 is known
- F distribution used when considering the ratio of two independent chi-square random variables χ_n^2 and χ_m^2

The probability distributions considered earlier in Module 2 arose from theoretical considerations

Note that the sampling distribution of a statistic will depend on –
(a) Distribution of the Population
(b) Sample Size and
(c) Sampling Strategy

Sampling Distributions of the sample mean $\bar{X}$ and the sample variance S^2.

If the random sample $X_1, X_2, \ldots\ldots, X_n$ is drawn from the normal distribution (i.e., $X_1, X_2, \ldots\ldots, X_n \sim N(\mu, \sigma^2)$) then by the linearity principle, we have $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

It turns out that $\bar{X}$ and S^2 are independent random variables where this independence of $\bar{X}$ and S^2 is a unique property of the normal distribution.

However, what is very interesting and useful is that we can state the asymptomatic distribution of $\bar{X}$ without assuming the distribution of the individual samples $X_1, X_2, \ldots\ldots, X_n$ (*Central Limit Theorem*)

Sampling Distribution of $\bar{X}$: ***Central Limit Theorem***

Central Limit Theorem: Consider a random sample of size n denoted by $X_1, X_2, \ldots\ldots, X_n$ which is drawn from a population with mean μ and variance σ^2 where $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Then the asymptotic distribution of $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is the standard normal distribution $N(0,1)$

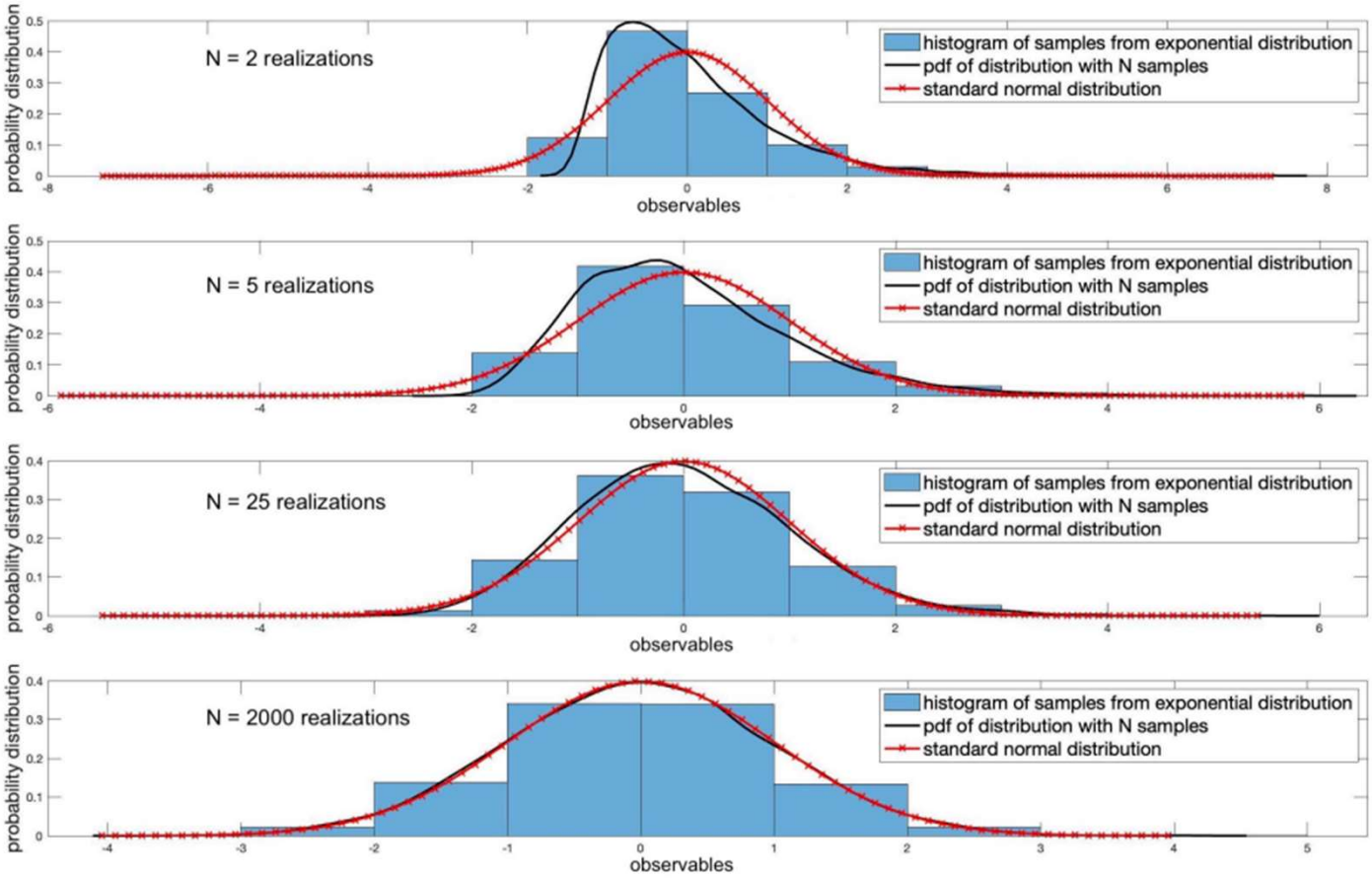
The important point here is that the **asymptotic distribution of $\boldsymbol{Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}$ for a large enough $\boldsymbol{n}$, is normally distributed, even though $\boldsymbol{X_1, X_2, \ldots\ldots, X_n}$ may not be normally distributed**

How large an n do we need for CLT to hold? How many samples do we need to take?

If the underlying population distribution is like a standard normal distribution, then only a few (3-5) samples would be enough. If the underlying distribution is very different from a *normal* one then we may need more, say ~ 30

Read pages 65-67 of Prof. Amrik's notes for details.

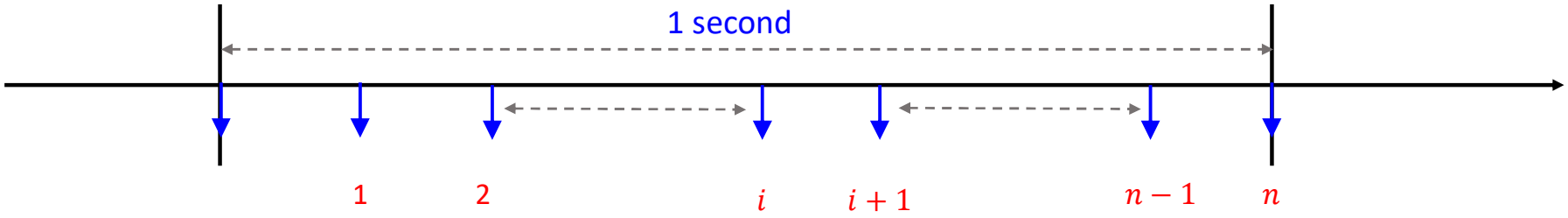
The Central Limit Theorem (CLT) demonstrates that as the size of the samples is increased from N = 2 to N = 2000, the probability distribution of the appropriately subtracted and normalized sample mean $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ (shown by black solid lines) resembles the standard normal distribution (shown by cross-cut red lines).



Example 6.4.4 Application of CLT (*please read the book for "camera" details*)

Distance measured n times a second. After each second, the mean distance measured is reported.

How large should n be to be at least 95% certain that the estimated distance is within $\pm 0.5m$ of the actual distance?



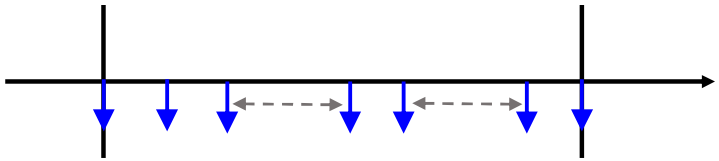
Distance Measurement Camera gives sample values $X_1, X_2, X_3, \ldots\ldots\ldots, X_n$, i.e., n samples every second, after which a decision is to be made.

Each of these $X_i s$ are i.i.d. random variables with mean $E(X_i) = \bar{X}_i = \mu$ and $Var(X_i) = E\left(\overline{X_i^2} - \bar{X}_i{}^2\right) = \sigma^2$

Note that we are given the mean $\mu\ (= d)$and the variance $\sigma^2 (= 4)$ of the random variables $X_i s$. We do not know their distribution, but we know/assume that they are i.i.d. random variables

$\bar{X}$ is the mean measurement made every second by averaging $X_1, X_2, X_3, \ldots\ldots\ldots, X_n$

$$\bar{X} = \frac{1}{n}\left[X_1 + X_2 + \ldots. + X_n\right]$$



Note that $\bar{X}$ is a random variable whose distribution we do not actually know but we claim from CLT that this will be a Normal Random Variable

From the measured values $X_1, X_2, X_3, \ldots\ldots\ldots, X_n$ we get the estimated mean $E(\bar{X})$ and variance $Var(\bar{X})$ as -

$$E\left(\bar{X}\right) = \frac{1}{n}\left[E\left(X_1\right) + E\left(X_2\right) + \ldots. + E\left(X_n\right)\right] = d$$

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n} = \frac{4}{n}$$

See Slide 12

From CLT,

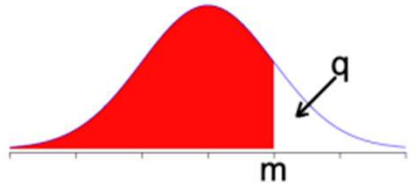
$$\bar{X} \sim N\left(d, \frac{4}{n}\right)$$

With Z as the standard normal r.v. with $mean = 0$ and $variance = 1$, we get -

$$P(-0.5 < \bar{X} - d < 0.5) = P\left(\frac{-0.5}{2/\sqrt{n}} < \frac{\bar{X} - d}{2/\sqrt{n}} < \frac{0.5}{2/\sqrt{n}}\right) = P\left(-\frac{\sqrt{n}}{4} < Z < \frac{\sqrt{n}}{4}\right)$$

$$= 2P\left(Z < \frac{\sqrt{n}}{4}\right) - 1$$

≥ 0.95 for at least 95% certainity



Another useful result for $N(0,1)$ distribution

For $q = 1 - P(Z < m), \quad m > 0$

$$P(-m < Z < m) = (1 - 2q)$$

$$= 1 - 2\{1 - P(Z < m)\}$$

$$= 2P(Z < m) - 1$$

Table of the Standard Normal Distribution (taken from your textbook)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
⋮										
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Figure 6.8: An excerpt from the standard normal distribution table compatible with the shaded portion of the probability distribution shown in Figure 6.7. In the example discussed in section 6.4.4, $m = 1.96$. $P(Z < 1.96)$ can be computed from the table by reading off the entry corresponding to $(1.9, 0.06) = 0.9750$.

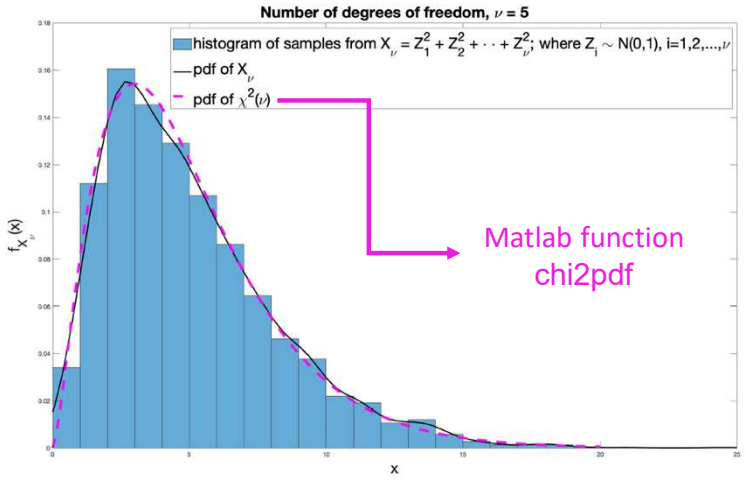
$$2P\left(Z < \frac{\sqrt{n}}{4}\right) - 1 \geq 0.95 \quad \Rightarrow \quad P\left(Z < \frac{\sqrt{n}}{4}\right) \geq 0.975 \qquad \frac{\sqrt{n}}{4} \geq 1.96 \quad \Rightarrow \quad n \geq 62$$

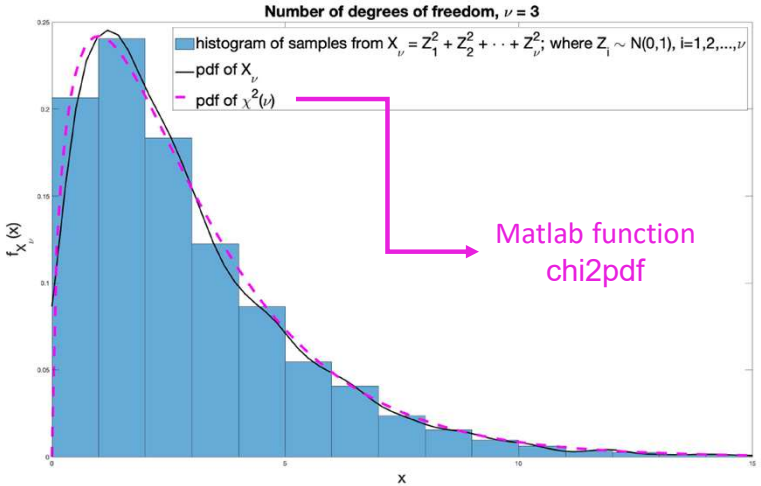
Study of some Sampling Distributions which either arise from the Standard Normal Distribution or Converge Asymptotically to the Standard Normal Distribution

(i) The chi-square distribution $X \sim \chi^2(\nu)$

This is the distribution of the *sum of the squares of ν independent standard normal random variables* $\mathcal{N}(0,1)$ and is said to have ***ν degrees of freedom***

$$X_\nu \sim \chi^2(\nu): \qquad X_\nu = Z_1^2 + Z_2^2 + \ldots\ldots\ldots + Z_\nu^2 \qquad Z_i \sim N(0,1); i = 1,2,\ldots\ldots,\nu$$





Note that $\chi^2(1)$ is the **square** of standard normal random variable $Z \sim N(0,1)$; see next slide!

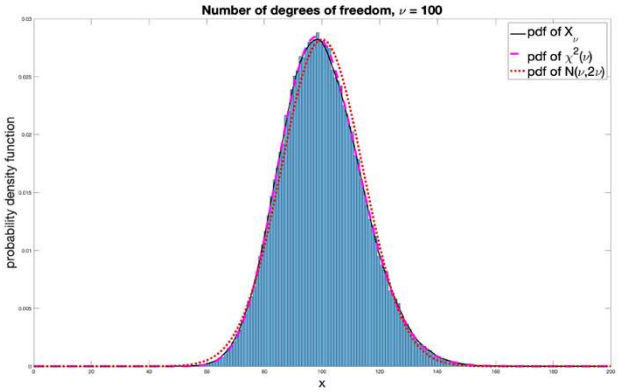
Properties of the chi-square random variables:

* χ^2 values are always positive (for $\nu > 1$) or non-negative (for $\nu = 1$)

* The shape of the pdf of χ^2 random variables depends on ν

* $E(X_\nu) = \nu$ and $Var(X_\nu) = 2\nu$

* $\chi^2(\nu) \to N(\nu, 2\nu)$ as $\nu \to \infty$



For large ν(say 100), the $\chi^2(\nu)$ distribution converges to $N(\nu, 2\nu)$

This approximation is very good for large values of ν, say for $\nu \geq 30$.

This implies that for large values of ν, $Z := \frac{X_\nu - \nu}{\sqrt{2\nu}} \sim N(0,1)$ where $X_\nu \equiv \chi^2_\nu$

- pdf of $\chi^2(n)$ $\quad f_X(x) = \dfrac{1}{2}\dfrac{\left(\frac{x}{2}\right)^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\Gamma\left(\frac{n}{2}\right)}, \quad x \geq 0$

Γ Gamma Function

$\chi^2(2)$: $f_X(x) = \dfrac{1}{2}e^{-\frac{x}{2}} \quad x \geq 0$

Exponential Distribution

Application of the χ^2 Distribution: Sampling Distribution of $\frac{(n-1)S^2}{\sigma^2}$

Let $X_i \sim N(\mu, \sigma^2); i = 1,2, \ldots, n$ be the n samples with

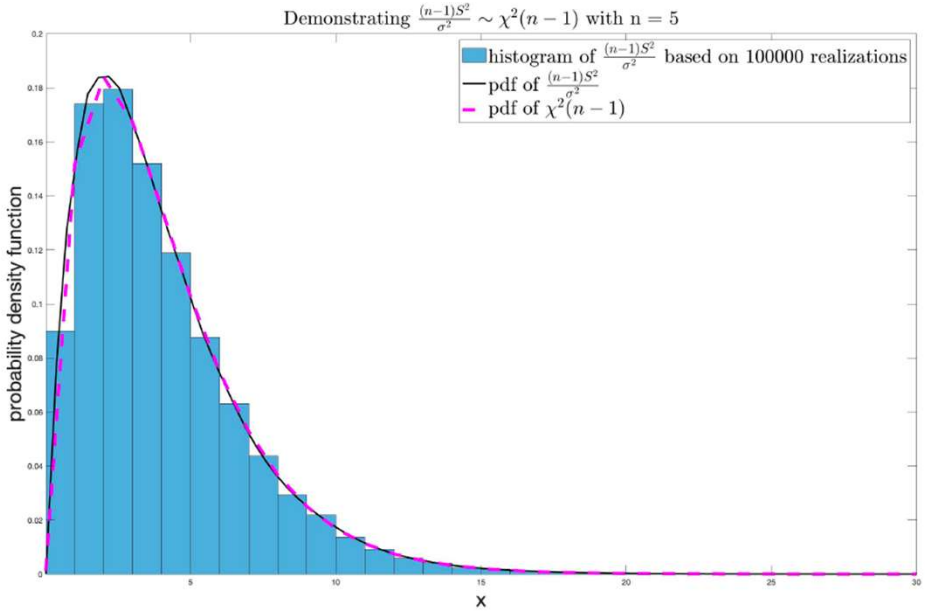
Sample Mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and

Unbiased Sample Variance $S^2 := \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Then, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

The r.v. $\frac{(n-1)S^2}{\sigma^2}$ has a χ^2 distribution with $(n-1)$ degrees of freedom, i.e., it has the distribution of the sum of squares of $(n-1)$ normally distributed $N(0,1)$ random variable each with zero mean and unit variance.

Why is this true even though the summation is from $i = 1$ *to* n?



Sampling Distribution of the Sample Variance

Example: The Success Rate of a Sniper

The optical scope on the marksman's rifle has a lack of precision in each of the horizontal and vertical coordinates that is normally distributed with mean 0 and variance of 4 sq. meters. What is his success rate to hit a target within a radius of 0.1 m?

Solution: Let $R_{err}^2 = X^2 + Y^2$ denote the square of the error to hit the target $X, Y \sim N(0,4)$. We scale X, Y to $Z_1 = X/2$ and $Z_2 = Y/2$ to have Z_1 and Z_2 as *standard normal random variables* with $Z_i \sim N(0,1), i = 1, 2$. Therefore,

$$P\left(R_{err}^2 < 0.01\right) = P\left(Z_1^2 + Z_2^2 < 0.01/4 = 0.0025\right)$$
$$= P\left(\chi_2^2 < 0.0025\right) = 1 - e^{-\frac{0.0025}{2}}$$
$$= 0.0012$$

χ_2^2 becomes an exponential distribution

Therefore, the success rate of the sniper in the given situation is only 0.12%

Example: Locating a target in three-dimensional space, and the three coordinate errors (in meters) of the point chosen are independent r.v.s with mean 0 and std. dev. 2. Find the probability that the distance between the point chosen and the target exceeds 3 meters.

If D is the distance, then $D^2 = X_1^2 + X_2^2 + X_3^2$ where X_i is the error in the i^{th} coordinate.

Since $Z_i = \frac{X_i}{2}, i = 1, 2, 3$ are all standard normal rvs, it follows that

$$\begin{aligned} P[D^2 > 9] &= P[X_1^2 + X_2^2 + X_3^2 > 9] = P\left[Z_1^2 + Z_2^2 + Z_3^2 > \frac{9}{4}\right] \\ &= P\left[\chi_3^2 > \frac{9}{4}\right] \\ &= 0.522 \qquad \textit{from tables or MATLAB} \end{aligned}$$

(ii) The t-Distribution (also known as the Student's t-Distribution) $T \sim t(n-1)$

Consider when we take n random samples Y_i from any population distribution which has the mean μ and variance σ^2, where both are known.

Then, using CLT, as $n \to \infty$, we can approximate the sample mean $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$\Rightarrow \quad Z := \frac{\bar{Y} - \mu}{\sqrt{\left(\frac{\sigma^2}{n}\right)}} \sim N(0,1)$$

Shifting the r.v. $\bar{Y}$ and then scaling it to get the r.v. Z

However, in many practical scenarios, Z may not be a suitable statistic to use as the variance σ^2 may not be known or known only very approximately (e.g., when the sample size is small.)

In that case, we can modify this to an alternative test statistic where the Sample Variance S^2 is used instead of σ^2. This is also convenient to do, as S^2 is knowable/computable.

Note that when there are only a few sample points, the CLT approximation may not be very good anyway, and it may be better to use the $t-$ Distribution instead.

We know that for the χ^2 distribution, we have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

This motivates a new test statistic T as $T \equiv T_{n-1} := \frac{Z}{\sqrt{\frac{\chi^2_{(n-1)}}{(n-1)}}}$

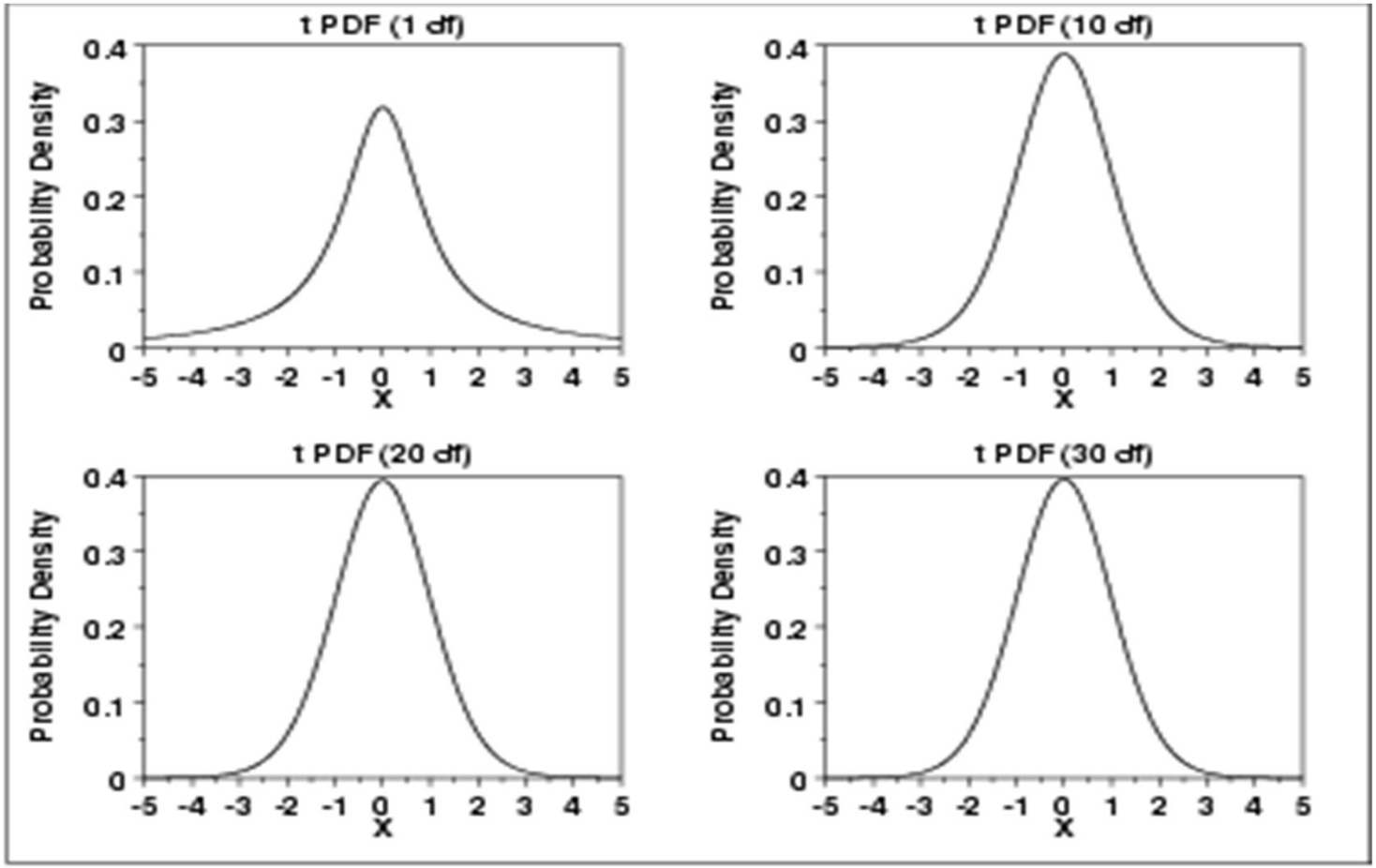
Note that here, both Z and $\chi^2_{(n-1)}$ are independent random variables

The test statistic T is formally given as –

$$T = \frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

t distribution with $(n-1)$ degrees of freedom

Easy to see that as $n \to \propto$, this tends towards a standard normal distribution $N(0,1) \Rightarrow$ When the number of samples is large, it may just be easier to use a normal distribution and still be accurate enough



PDF of $T \sim t(n-1)$ for different n

Properties of the t Distribution

1. $t(n) \to N(0,1)$ for $n \to \infty$

2. $E(T) = 0$ for $n > 1$ (otherwise, undefined) and $Var(T) = \frac{n}{n-2}$ for n>2

3. For α, 0< α<1, for the t-distribution with n degrees of freedom, let $t_{\alpha,n}$ be such that

$$P\{T_n \geq t_{\alpha,n}\} = \alpha \qquad \text{(A)}$$

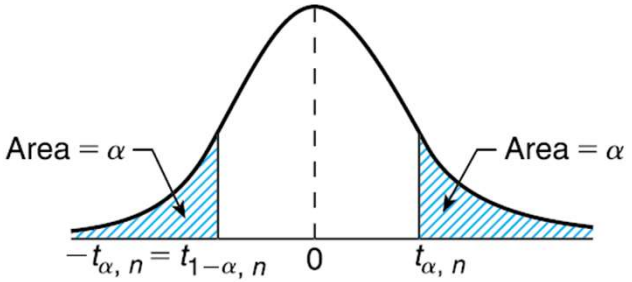
Then, it follows from the symmetry of the t-distribution about zero, that $-T_n$ has the same distribution as T_n .

Therefore -

$$\alpha = P\left\{-T_n \geq t_{\alpha,n}\right\} = P\left\{T_n \leq -t_{\alpha,n}\right\} = 1 - \underbrace{P\left\{T_n \geq -t_{\alpha,n}\right\}}_{=\alpha}$$

$$\Rightarrow \quad P\left\{T_n \geq -t_{\alpha,n}\right\} = 1 - \alpha$$

comparing with (A) $\Rightarrow$ $-t_{\alpha,n} = t_{1-\alpha,n}$



Typical values of α=0.05, or 0.01

Applications of the t Distribution

Consider two different samples of **sizes n_1 and n_2** with respective **means $\bar{x}_1$ and $\bar{x}_2$** and **sample variances S_1 and S_2**

Question: Are the two means sufficiently alike to infer that both samples were drawn from the same population?

To answer this, suggest the test statistic

$$T := \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{(n_1 - 1) + (n_2 - 1)}}{\sqrt{S_1 + S_2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$ with the t-Distribution

For a more detailed description of this, read Sheldon Ross (2021) (reference text) pp 326-338 or the book by R.A. Fisher (1925) pp 90-104 referenced in your textbook.

Example: Estimating the Spread of Viral Infection

1. The daily number $\boldsymbol{X}$ of reported flu cases is $X \sim N(70,9)$.

What is the probability that on a given day, the total number of reported cases exceeds 75?

$X \sim N(70,9) \quad \Rightarrow \quad Z := \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{75-7}{3} \approx 1.67,$

Z is a standard $N(0,1)$ normal random variable

Note n=1 in this case as there is only one report to be considered.

In terms of the original random variable X, we want to find $P(X > 75)$

Equivalently, in terms of the random variable Z, we want to find $P(Z > 1.67)$

We get $P(Z > 1.67) = 1 - P(Z \leq 1.67) = 0.0475$ (from MATLAB, or probability tables)

P(total number of reported cases in a day >75) = 0.0475

Example: Estimating the Spread of Viral Infection

2. The actual daily mean number of flu cases is 70 (i.e. , $\mu_Y = 70$) but it is not known if Y follows a normal distribution. Over 15 days, the sample mean $\bar{Y}$ of the number of infections is computed. It is also observed that the Sample Standard Deviation is $S = 4$.

Question: What is the probability $P(\bar{Y}>74)$?

Since the population variance is unknown, the test statistic used is $T = \frac{\bar{Y}-\mu_Y}{\frac{S}{\sqrt{n}}} \sim t(n-1)$

as it depends on the sample standard deviation $S = 4$. Therefore,

$$P\left(\bar{Y} > 74\right) = P\left(\frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}} > \frac{74-70}{\frac{4}{\sqrt{15}}}\right) = P\left(T > 3.8730\right) = 0.00084461$$

$T \sim t(n-1)$ *used with* $n = 15$

The F Distribution $\sim F(n,m)$ with n and m degrees of freedom

If $X_n = \chi^2_n$ and $X_m = \chi^2_m$ are two independent chi-square random variables with n and m degrees of freedom, respectively, then the ratio

$$F_{n,m} := \frac{X_n/n}{X_m/m}$$

is a random variable with the F distribution with n and m degrees of freedom.

The corresponding pdf is -

$$f_X(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \quad 0 < x < \infty$$

The F Distribution

Further, if we have two independent samples of size n_1 and n_2 from two independent normal populations with respective variances σ_1^2 and σ_2^2, then the statistic

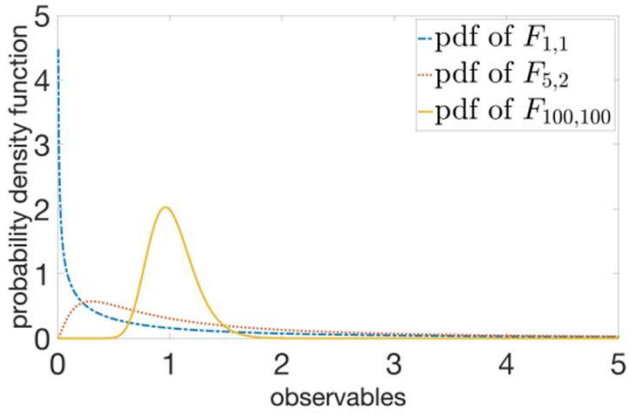
$$F := \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

S_1 = Sample Variance of the n_1 samples
S_2 = Sample Variance of the n_2 samples

follows the $F(n_1 - 1, n_2 - 1)$ distribution

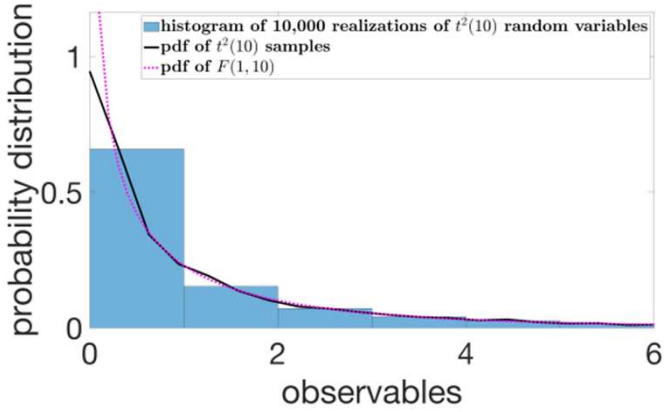
Properties of the F Distribution

1. The F distribution is not defined for negative values
2. The pdf of the F random variables is not symmetric in shape over the range of the observables



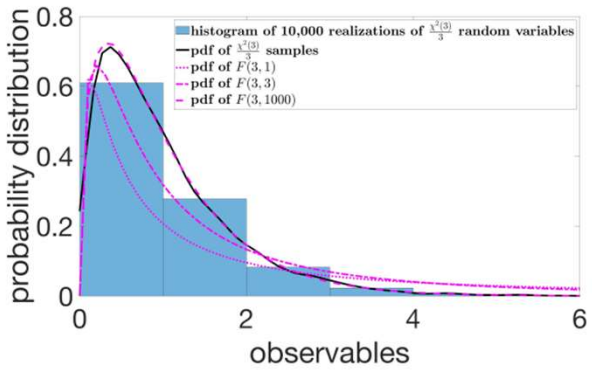
3. $F(1,m) \equiv t^2(m)$

$$t(m) = \frac{Z}{\sqrt{\chi_m^2 / m}} = \frac{\sqrt{\chi_1^2 / 1}}{\sqrt{\chi_m^2 / m}} = \sqrt{F_{1.m}}$$



4. $F(n,m) \to \frac{\chi_n^2}{n}$ as $m \to \infty$

$$\lim_{m\to\infty} \frac{\chi_m^2}{m} = 1$$



Properties of the F Distribution .....*continued*.....

5. The Mean and Variance of the $\sim F(n,m)$ distribution, i.e., $F := \frac{X_n/n}{X_m/m}$

Mean $E(F) = \frac{m}{m-2}$ Variance $Var(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$

6. Consider $P(F > f_{\alpha,n,m}) = \int_{f_{\alpha,n,m}}^{\infty} f_F(x)dx = \alpha$

$P(F \leq f_{\alpha,n,m}) = \int_0^{f_{\alpha,n,m}} f_F(x)dx = 1 - \alpha$

$P(F \leq f_{1-\alpha,n,m}) = \int_0^{f_{1-\alpha,n,m}} f_F(x)dx = \alpha$

Here $f_{\alpha,n,m}$ is the *upper-tailed α-percentage point*

and $f_{1-\alpha,n,m}$ is the *lower-tailed $(1-\alpha)$-percentage point*

α corresponds to the *rejection region*.
The area under the pdf $f_F(x)$ to the right of $f_{\alpha,n,m}$ is equal to α.

(1- α) corresponds to the *acceptance region,* the area under the pdf $f_F(x)$ to the left of $f_{\alpha,n,m}$

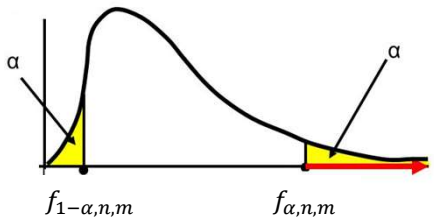
The lower and the upper points are related as

$$f_{1-\alpha,n,m} = (f_{\alpha,m,n})^{-1}$$

See next slide for details

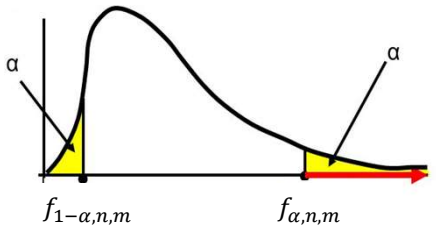
Here α is known as the *level of significance*

and $f_{\alpha,n,m}$ as the *critical value*



If $f_{\alpha,n,m}$ is the upper-tailed α-percentage point for this distribution $W \sim F_{n,m}$, then

$$P(W > f_{\alpha,n,m}) = \alpha$$



Choose $f_{1-\alpha,n,m}$ such that $$\alpha = P\left[W \le f_{1-\alpha,n,m}\right] = P\left[\frac{1}{W} \ge \frac{1}{f_{1-\alpha,n,m}}\right]$$

For a lower-tailed probability of α, where α is generally small, we note that if the distribution of W is $W \sim F_{n,m}$, then the distribution of $1/W$ is $F_{m,n}$.

Therefore, $$P\left[\frac{1}{W} \ge f_{\alpha,m,n}\right] = \alpha$$

It then follows that $$\frac{1}{f_{1-\alpha,n,m}} = f_{\alpha,m,n}$$

Applications of the F Distribution

1. The test statistic used for performing an *analysis of variance* (ANOVA) experiment to test the difference between means of different populations is an F random variable that follows the F distribution.

2. The F distribution is also used to test the existence of any significant difference between the variances of two different groups of population.

   For example, a university academic policy may prefer that two instructors, co-teaching a course, should grade exams in such a way so as to have the same variation in their grading

Hypothesis Testing & Statistical Inference

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H_0.

An alternative hypothesis (denoted H_1 or H_a), which is the opposite of what is stated in the null hypothesis, is then defined.

The hypothesis-testing procedure involves using sample data to determine whether or not H_0 can be rejected. If H_0 is rejected, the statistical conclusion is that the alternative hypothesis H_1 is true.

Components of an experiment to test hypothesis:

The key components are –

1. Construct the statement to be tested: $null\ (H_0) vs. alternate\ (H_1\ or\ H_a)$ hypothesis
2. Identify the *rejection* (or *critical*) region to enable a decision about the hypothesis.
   For example, the evidence based on the test statistic may suggest that we either *reject* or *fail to reject* the null hypothesis.
3. Quantify the likely error in the decision arrived at in (2) in terms of a probability measure. This could be either a –

$type-1\ error$ mistaken rejection of an actually *true null hypothesis* with a probability of occurrence α

or $type-2\ error$ mistaken acceptance of an actually *false null hypothesis* with a probability of occurrence β

See next slide

Our objective will generally be to reduce these errors while deciding, but in many scenarios, reducing one type of error can increase the other type of error

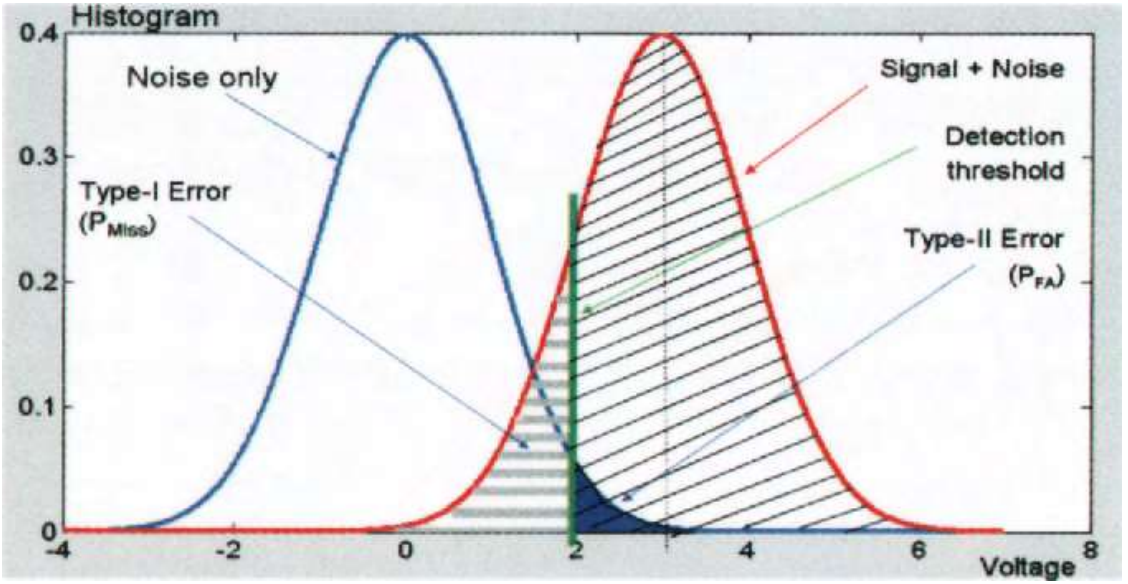
Example: Detecting a Reflected Radar Signal from an Aircraft in a Noisy Environment

Return Signal
(from an aircraft)

+

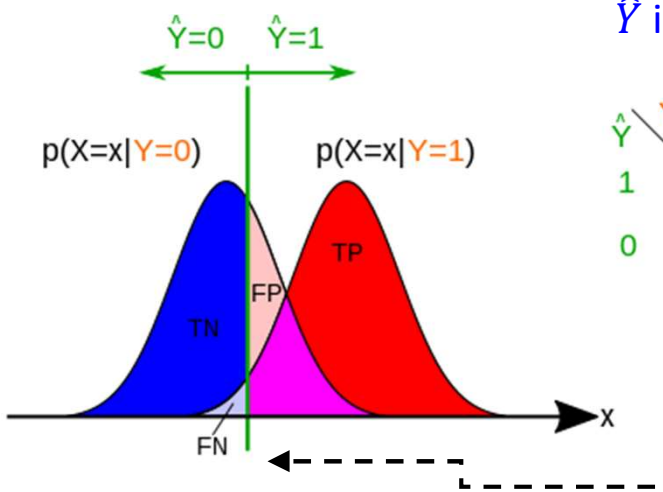
Ambient Noise

H_0: Enemy Aircraft Present
H_1: No Enemy Aircraft Present



From: https://www.semanticscholar.org/

Type - I Error: Missed Detection — Signal present but (Signal + Noise) < Detector Threshold
Type -II Error: False Alarm — No Signal but Noise > Detector Threshold



$\hat{Y}$ is the estimate of Y

$\hat{Y}$ \ Y	1	0
1	TP	FP
0	FN	TN

Moving the threshold left or right forces a tradeoff between the two types of error!

TRUE POSITIVE (TP: 1, 1) Correct Decision
TRUE NEGATIVE (TN: 0:0) Correct Decision

Type 2 Error Probability β
FALSE POSITIVE (FP: 1, 0) Wrong Decision
{0 interpreted as 1}

Type 1 Error Probability α
FALSE NEGATIVE (FN: 0, 1) Wrong Decision
{1 interpreted as 0}

Decision \Truth condition	H_0 is true	H_0 is not true
H_0 is not rejected	Decision is correct (with probability $1-\alpha$)	type-2 error (with probability β)
H_0 is rejected	type-1 error (with probability α)	Decision is correct (with probability $1-\beta$)

Steps for Performing a Hypothesis Test

1. Specify H_0 and H_1 and an acceptable level of significance α

2. Define a sample-based test statistic (e.g., $\bar{X}, S^2$ etc.) and a rejection (or critical) region for H_0 that is most suitable for the experiment.

3. Collect the sample data and calculate the test statistic

4. Make a decision to either *reject or fail to reject* H_0

5. Interpret the results in the language of the of the problem at hand (e.g., provide confidence intervals, etc.) and provide an estimate of the error in the decision.

What might determine our choice of α?

Example (from the textbook, page 82)

A company that packages salted peanuts in 8 kg jars is interested in maintaining control on the number of peanuts put in the jars by one of the machines in its packaging units.

Control is defined *as averaging 8 kg per jar* and *not consistently over or under filling the jars*. To monitor this control, a sample of 16 jars is taken from the packaging line at random time intervals and their contents weighed. The mean weight of peanuts in these 16 jars will be used to test the null hypothesis that the machine is indeed working properly.

What may be a suitable **level of significance α** for this test?

For convenience, we assume that the population standard deviation $\sigma = 0.2$ kg of the weight of the jars is known to us.

Choosing α for the previous example – $(\mu = 8, \sigma = 0.2, N = 16)$

In this case, we are only interested in H_0 and the corresponding level of significance α

1. (Step 1) Choose the hypothesis as - $H_0: \mu = 8$ $H_a: \mu \neq 8$

2. (Steps 2 & 3) Choose the test statistic for this case as the Sample Mean $\bar{X} = \frac{1}{16}\sum_1^{16} X_i$

3. (Step 4) Suitable *rejection criteria* is then selected, such as $\bar{X} < 7.9$ or $\bar{X} > 8.1$

4. (Step 5) This helps in deciding the level of significance α of the test estimated as follows - (This may also be interpreted as the *maximum Type-1 error.*)

H_a could have been –
$\mu > 8$ overfilling
$\mu < 8$ underfilling

$$\alpha = P(\bar{X} < 7.9 \text{ or } \bar{X} > 8.1 \text{ when } \mu = 8)$$

$$P(\bar{X} < 7.9) = P\left(Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{7.9-8}{0.2/\sqrt{16}} = -2.0\right) = P(Z < -2.0)$$

$$= 0.0228$$ (e.g., from tables such as the one in the next slide)

Similarly (from symmetry about the mean) $P(\bar{X} > 8.1) = P(Z > 2.0) = 0.0228$

Therefore, α = P(Type-1 Error)= $P(\bar{X} < 7.9) + P(\bar{X} > 8.1)$ =0.0456 Note that $\bar{X} < 7.9$, $\bar{X} > 8.1$ are disjoint

Type 1 error: The null hypothesis is true (i.e., item is within the "acceptance range") but we make the error of rejecting it. The Level of Significance α is then the probability of that happening (i.e., the probability of a type-1 error)

Table 1: Table of the Standard Normal Cumulative Distribution Function $\Phi(z)$

P(z < -2) = .0228

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

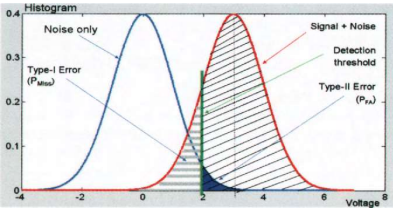
Comments on the Level of Significance α

1. There is no general rule of thumb to choose α as we may not have a clear idea of what an appropriate maximum allowable *type-1 error* should be for a typical statistical experiment.

2. The level of significance α may also be sensitive to minor changes in the sample statistic which will affect testing the veracity of the hypothesis appropriately

3. There is always a trade-off between α (probability of *type 1 error*) and β (probability of *type-2 error*) as any effort to reduce one may be likely to increase the other. *Jars are being filled all right, but we say they are not* vs *Jars are not being filled all right, but we say they are*.

   (See also the earlier slide on *missed detection* and *false alarm.)*
   Changing the detection threshold to reduce probability of false alarm will increase the probability of missed detection



Two Sample Test for Means: Another test of hypothesis using the t distribution

Consider a case where we have two different populations that are normally distributed with the same variance. A random variable, sampled from each population, is denoted by $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$

Let there be n_1 samples taken from the first population, i.e., $X_{1i} \sim N(\mu_1, \sigma^2)$; $\quad i = 1, 2, \ldots\ldots, n_1$
and n_2 samples taken from the second population, i.e., $X_{2j} \sim N(\mu_2, \sigma^2)$; $\quad j = 1, 2, \ldots\ldots, n_2$

1. Step 1: Construct the hypothesis H_0: $\mu_1 = \mu_2$
and H_1: $\mu_1 \neq \mu_2$ double sided test
(alternatively, $\mu_1 > \mu_2$ ***or*** $\mu_1 < \mu_2$) single sided test
Further, choose and set α

2. Steps 2 &3: The test statistic is $t := \frac{(\overline{X_1} - \mu_1) - (\overline{X_2} - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$

and S_j^2 is the sample variance corresponding to the samples taken from the j^{th} population set $j = 1,2$

3. Step 4: Identify the rejection criteria. There may be three distinct cases depending on the type of test (double or single-sided alternate hypothesis)

- Reject H_0 in favour of H_1 $(\mu_1 \neq \mu_2)$ if

$$|t| \geq t\left(\frac{\alpha}{2}, (n_1 - 1) + (n_2 - 1)\right)$$

- Reject H_0 in favour of H_1 $(\mu_1 > \mu_2)$ if

$$t \geq t\big(\alpha, (n_1 - 1) + (n_2 - 1)\big)$$

- Reject H_0 in favour of H_1 $(\mu_1 < \mu_2)$ if

$$t \geq -t\big(\alpha, (n_1 - 1) + (n_2 - 1)\big)$$

Note that in the above, the RHS term refers to the t observable value from the t distribution with a given significance level ($\frac{\alpha}{2}$ or α, as stated above) and $\big((n_1 - 1) + (n_2 - 1)\big)$ degrees of freedom.

Example (6.6.8): Choosing between two gasoline brands for optimal mileage and performance

While comparing two different gasoline brands, a consumer survey reveals the following:

- A full tank of brand **Gusto** requires 4 cans and covers 546 km with a standard deviation of 31 km
- A full tank of brand **Jiva** requires 4 cans and covers 492 km with a standard deviation of 26 km

Assume that the performance parameters (mentioned above) of both brands are sampled from Normal distributions with equal variances.

Test if there is a significantly better value in terms of mileage offered by Gusto over Jiva or if the mileage of both brands are statistically similar. Choose $\alpha = 0.05$.

Solution: In the following, we use the subscript G for brand Gusto and subscript J for brand Jiva and define the hypothesis as follows.

H_0: $\mu_G = \mu_J$ $\quad$ H_1: $\mu_G > \mu_J$

From samples of **Gusto**, we have $\overline{X_G} = 546, S_G = 31, n_G = 4$
From samples of **Jiva**, we have $\overline{X_J} = 492, S_J = 26, \ n_J = 4$

$$S^2 = \frac{(n_G - 1)S_G^2 + (n_J - 1)S_J^2}{(n_G - 1) - (n_J - 1)}$$

$$t := \frac{(\overline{X_G} - \mu_G) - (\overline{X_J} - \mu_J)}{S\sqrt{\frac{1}{n_G} + \frac{1}{n_J}}}$$

The sample variance $S^2 = \frac{(4-1)31^2 + (4-1)26^2}{4+4-2} \Rightarrow S = 28.609$

Therefore, $t\ (under\ H_0) = \frac{54}{28.609\sqrt{\frac{1}{4}+\frac{1}{4}}} = 2.67$

The observable value $t(0.05,6) = 1.9432$ (either from the t distribution look-up table or using Matlab) where $(n_G - 1 + n_J - 1) = 4 + 4 - 2 = 6$.

Since this is a single-tailed test H_1: $\mu_G > \mu_J$ and because $t_{calculated} = 2.67 > t(0.05,6) = 1.9432$ we reject H_0 in favour of H_1.

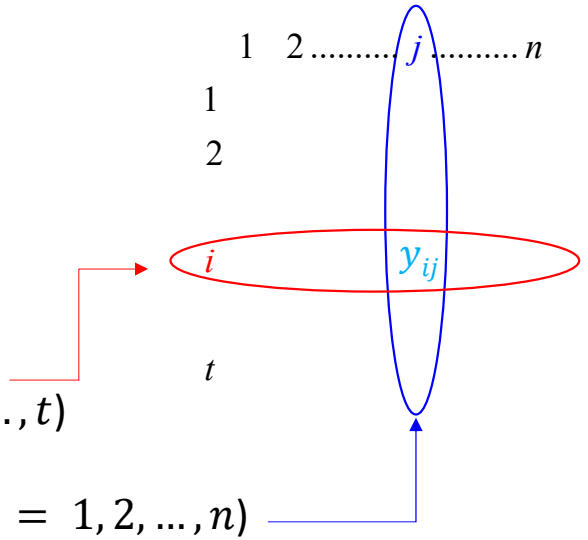
Reject H_0 in favour of H_1 $(\mu_G > \mu_J)$ if $t \geq t\left(\alpha, (n_G - 1) + (n_J - 1)\right)$

In other words, brand Gusto will likely give us better mileage than brand Jiva at the level of significance $\alpha = 0.05$ (i.e., with 95% confidence level)

The Two Sample Test for Means cannot be generalized for more than two different population sets.

For multi-population tests, we may have to resort to *Analysis of Variance (ANOVA)* test which are described next

One-way ANOVA



Data collated from survey samples is denoted by y_{ij} where the,

- first subscript represents the i^{th} population group ($i = 1, 2, \dots, t$)

and - second subscript represents the j^{th} observation (data point; $j = 1, 2, \dots, n$) in the group

We will consider $n_1, n_2, \dots, n_t$ observations for each of the t population groups.

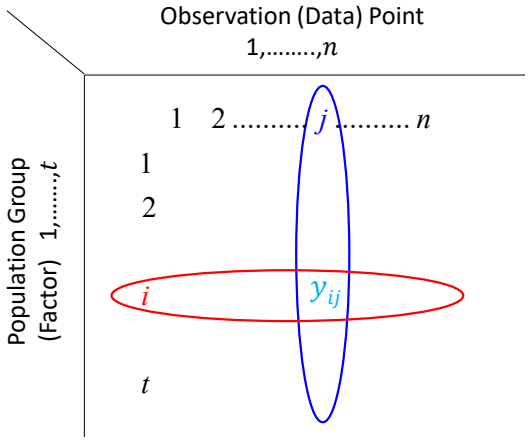
For the special case where $n_1 = n_2 = \dots = n_t = n$, we have a *balanced data set*

The total number of observations is $\sum_{i=1}^{t} n_i$ ($= nt$ in the case of balanced data)

Null Hypothesis (One Way ANOVA)

H_0: $\mu_1 = \mu_2 \ldots\ldots = \mu_t$

H_1: the above equalities are not satisfied



Assumption:

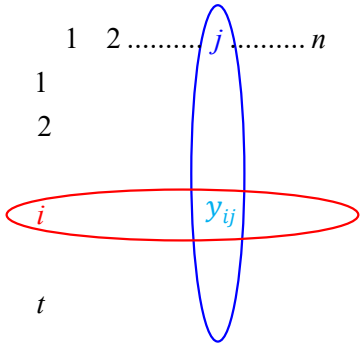
Data in each of the t population groups is normally distributed as $N(\mu_i, \sigma^2)$ $i = 1, \ldots, t$ where the variance σ^2 is the same across all the population groups

Approximation appealing to the CLT is possible

nt total observation points with n observation points in each of the t groups

One-Way ANOVA can also be formulated for scenarios where the number of observation points in the t groups are not same.
(See Ross for other basic variations)

What is the approach of (One-Way) ANOVA?

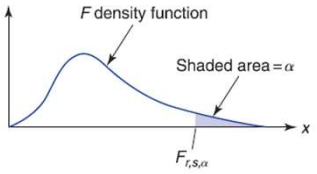


Let $\bar{Y}_i$ and S_i^2 be the sample mean and sample variance of the data of the i^{th} population group $i = 1, \dots, t$.

We test H_0 by comparing the values of two estimators of the common variance σ^2

First Estimator: Estimates σ^2 only when H_0 is true; **tends to exceed σ^2 when H_0 is not true**

Second Estimator: Estimates σ^2 directly and is always valid, whether or not H_0 is true



Random variable F with degrees of freedom r, s: $P\{F \geq F_{r,s,a}\} = \alpha$.

Define a suitable F statistic as the ratio $\frac{First\ Estimator}{Second\ Estimator}$ of the above two (independent) estimates of the common variance σ^2.

Ratios of two chi-sq r.v. will have the F-distribution

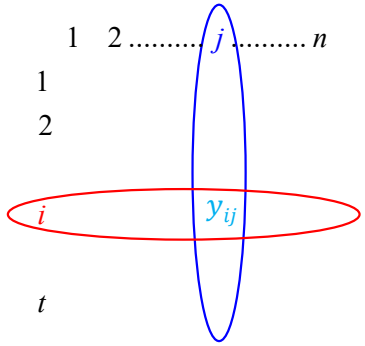
Calculate the probability of this statistic ratio **exceeding** some acceptable threshold to reject H_0

Values of $F_{r,s,0.05}$

s = Degrees of freedom for denominator	r = Degrees of freedom for numerator			
	1	2	3	4
4	7.71	6.94	6.59	6.39
5	6.61	5.79	5.41	5.19
⋮	⋮	⋮	⋮	⋮
10	4.96	4.10	3.71	3.48

Estimator of Variance not dependent on H_0

There are tn independent observations (i.i.d. normal r.v.s) $y_{ij} \sim N(\mu_i, \sigma^2)$ $i = 1, \ldots, t$ $j = 1, \ldots\ldots, n$ where both the mean μ_i and the variance σ^2 are unknown.



It follows that $$\sum_{i=1}^{t}\sum_{j=1}^{n}\frac{\left(y_{ij}-E[y_{ij}]\right)^2}{\sigma^2}=\sum_{i=1}^{t}\sum_{j=1}^{n}\frac{\left(y_{ij}-\mu_i\right)^2}{\sigma^2}\sim\chi^2_{tn}$$ **[A]**

Has a chi-square distribution with tn degrees of freedom
μ_i is the true mean of the i^{th} population, not known

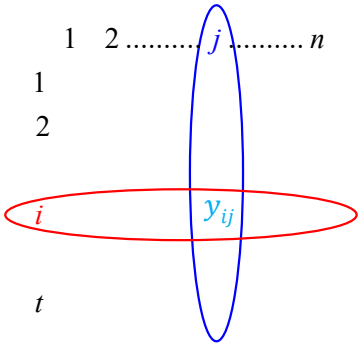
To obtain estimators for the t unknown parameters μ_1 ..... μ_t, we use $\bar{y}_{i\,.}$ to denote the row average $\bar{y}_{i\,.}=\sum_{j=1}^{n}\frac{y_{ij}}{n}$

This $\bar{y}_{i\,.}$ is then the sample mean of the i^{th} population and is an *estimator* of the mean μ_i $i = 1, .., t$

Using this for μ_i in [A], we can see that $\sum_{i=1}^{t}\sum_{j=1}^{n}\frac{\left(y_{ij}-\bar{y}_{i.}\right)^2}{\sigma^2}\sim\chi^2_{tn-t}$ (chi-sq with $tn - t$ degrees of freedom since one degree of freedom is lost for each of the estimated parameters)

Note also that the mean of this chi-squared random variable will be $(tn - t)$

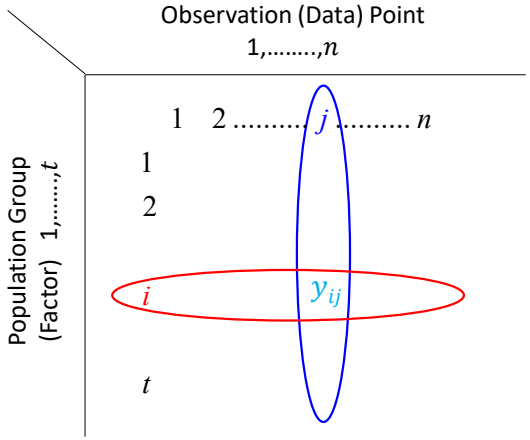
Let $SS_P = \sum_{i=1}^{t}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$ and $E[SS_P] = (tn - t)\sigma^2$



Estimator of Variance not dependent on H_0 $s_p^2 = \dfrac{SS_P}{\sum_{i=1}^{t}(n_i - 1)}$ $\dfrac{SS_p}{nt-t}$ for the balanced case

Note once again that this estimator was obtained without assuming anything about the truth of the null hypothesis H_0

Estimator of Variance when H_0 is TRUE $\Rightarrow$



- All the population means are equal $\mu_i = \mu \quad i = 1, \ldots, t$
- The m sample means $\bar{y}_{1.}$, ..., $\bar{y}_{t.}$ are $\sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and are independent random variables
- Therefore, the sum of the squares of the t standardized random variables $\frac{\bar{y}_{i.}-\mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\,\frac{(\bar{y}_{i.}-\mu)}{\sigma}$ will be a chi-square random variable with t degrees of freedom
- Therefore, when H_0 is TRUE $\quad n\sum_{i=1}^{t}\frac{(\bar{y}_{i.}-\mu)^2}{\sigma^2} \sim \chi_t^2$
- The estimator of μ is $\bar{y}_{..} = \frac{\sum_{i=1}^{t}\bar{y}_{i.}}{t}$. Using this in the earlier expression in the line above, we get that, when H_0 is TRUE, we have $n\sum_{i=1}^{t}\frac{(\bar{y}_{i.}-\bar{y}_{..})^2}{\sigma^2} \sim \chi_{t-1}^2$ (lost one degree of freedom because of using the estimated mean)
- Define $s_{means}^2 = \frac{\sum_{i=1}^{t}(\bar{y}_{i.}-\bar{y}_{..})^2}{t-1}$ as the sample variance
- Note that when H_0 is TRUE, taking expectations, we get

$$n\frac{(t-1)s_{means}^2}{\sigma^2} = (t-1) = \text{mean of } \chi_{t-1}^2 \text{ r.v.}$$

Estimator of Variance when H_0 is TRUE is given by –

$$\sigma^2 = ns_{means}^2$$

Show that when H_0 is TRUE ,

$$\left[\frac{\sum_{i=1}^{t}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2}{(t-1)}\right] \geq \frac{\sigma^2}{n}$$

The equality holds only when H_0 is TRUE

Let $\mu_{.} = \frac{1}{t}\sum_{i=1}^{t}\mu_i$ be the average of the (true) means of the t population groups

For each population group $i = 1, \ldots, m$, let $Z_i = \bar{y}_{i.} - \mu_i + \mu_{.}$ or $\bar{y}_{i.} = Z_i + \mu_i - \mu_{.}$

Since $\bar{y}_{i.}$ is normal with mean $\mu_{.}$ and variance $\frac{\sigma^2}{n}$, Z_i would also be normal with mean $\mu_{.}$ and variance $\frac{\sigma^2}{n}$ for $i = 1, \ldots, m$

Let $Z_{.} = \sum_{i=1}^{t}\frac{Z_i}{t} = \bar{y}_{..} - \mu_{.} + \mu_{.} = \bar{y}_{..}$ or $\bar{y}_{..} = Z_{.}$

$\Rightarrow$ $\bar{y}_{i.} - \bar{y}_{..} = Z_i + \mu_i - \mu_{.} - Z_{.}$

Since $\bar{y}_{i.} - \bar{y}_{..} = Z_i + \mu_i - \mu_{.} - Z_{.}$

$$\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\left[(Z_i - Z_{.}) + (\mu_i - \mu_{.})\right]^2$$

$$= \sum_{i=1}^{t}(Z_i - Z_{.})^2 + \sum_{i=1}^{t}(\mu_i - \mu_{.})^2 + 2\sum_{i=1}^{t}(\mu_i - \mu_{.})(Z_i - Z_{.})$$

$$= (t-1)\frac{\sigma^2}{n} + \sum_{i=1}^{t}(\mu_i - \mu_{.})^2$$

Dividing by $(t-1)$, we get -

$$\left[\frac{\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2}{(t-1)}\right] = \frac{\sigma^2}{n} + \underbrace{\frac{\sum_{i=1}^{t}(\mu_i - \mu_{.})^2}{(t-1)}}$$

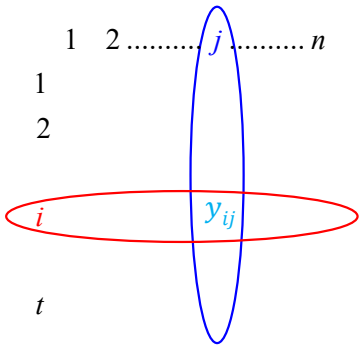
≥ 0

Zero only when

H_0 is TRUE

Organization of the Data (One Way ANOVA)

Population group	observations/data	$\sum_j y_{ij}$ (totals)	$\frac{Y_{i.}}{n_i}$ (means)	sum of squares
1	$y_{11}\ y_{12}\ \cdots\ y_{1n_1}$	$Y_{1.}$	$\bar{y}_{1.}$	SS_1
2	$y_{21}\ y_{22}\ \cdots\ y_{2n_2}$	$Y_{2.}$	$\bar{y}_{2.}$	SS_2
3	$y_{31}\ y_{32}\ \cdots\ y_{3n_3}$	$Y_{3.}$	$\bar{y}_{3.}$	SS_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
t	$y_{t1}\ y_{t2}\ \cdots\ y_{tn_t}$	$Y_{t.}$	$\bar{y}_{t.}$	SS_t
	Overall	$Y_{..}$	$\bar{y}_{..}$	SS_p



Dot Convention:

We have used the convention whereby the position of the . (dot) in the subscript represents which of the two indices (in the subscript) are being summed. For example, for some variable α_{ij}, we will use the following convention for summation:

$$\sum_i \alpha_{ij} = \alpha_{.j}$$

where the summation is performed over the first index i

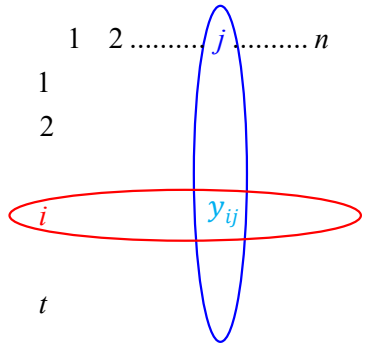
Sum of Squares: $SS_i = \sum_j (y_{ij} - \bar{\boldsymbol{y}}_{\boldsymbol{i}.})^2 \equiv \sum_j y_{ij}^2 - \frac{\boldsymbol{Y_{i.}^2}}{n_i}$

Pooled Sum of Squares: $SS_p = \sum_{i=1}^{t} SS_i$

Pooled Degrees of Freedom: $\sum_{i=1}^{t} n_i - t \ \ \{= (n-1)t\}$ *last equality holds for balanced data*

The Pooled Variance s_p^2 for One Way ANOVA is then defined as –

$$s_p^2 = \frac{SS_p}{\sum_{i=1}^{t}(n_i - 1)}$$



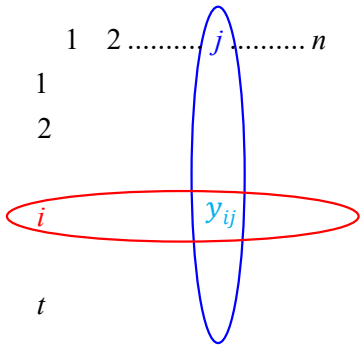
and for balanced data, we have –

$$s_p^2 = \frac{SS_p}{nt - t}$$

Note that this approach to estimating the variance has not used the null hypothesis at all.
It is also possible to do this estimation using the null hypothesis! (See next slide)

Alternative Approach for Estimating the Sample Variance (One Way ANOVA)

Population group	observations/data	$\sum_j y_{ij}$ (totals)	$\frac{Y_{i.}}{n_i}$ (means)	sum of squares
1	$y_{11}\ y_{12}\ \cdots\ y_{1n_1}$	$Y_{1.}$	$\overline{y}_{1.}$	SS_1
2	$y_{21}\ y_{22}\ \cdots\ y_{2n_2}$	$Y_{2.}$	$\overline{y}_{2.}$	SS_2
3	$y_{31}\ y_{32}\ \cdots\ y_{3n_3}$	$Y_{3.}$	$\overline{y}_{3.}$	SS_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
t	$y_{t1}\ y_{t2}\ \cdots\ y_{tn_t}$	$Y_{t.}$	$\overline{y}_{t.}$	SS_t
	Overall	$Y_{..}$	$\overline{y}_{..}$	SS_p



$$s^2_{means} = \frac{\sum_i (\overline{y}_{i.} - \overline{y}_{..})^2}{t-1}$$

Under the null hypothesis and based on the earlier discussions on sampling distributions, we may deduce that the factor level (*horizontal value*) means have a distribution with mean μ and variance $\frac{\sigma^2}{n}$.

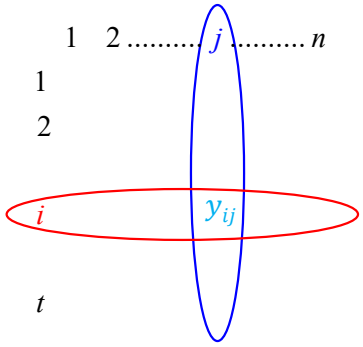
Therefore, we have an estimate for the population variance as $\sigma^2 = ns^2_{means}$ with $(t-1)$ degrees of freedom.

Of course, an alternate estimate of the population variance is s_p^2 with $t(n-1)$ degrees of freedom.

We know from the definition of the F statistic that the F *value represents the ratio of two independent estimates of a common variance*.

Therefore, $$F_{cal} = \frac{ns^2_{means}}{s_p^2}$$

If $F_{cal} > F_\alpha(t-1, t(n-1))$ then we reject H_0



Alternative Formulation for ANOVA (leads to the same inference)

SSB (sum of sqs. between groups) = $\sum_i \frac{(Y_{i.})^2}{n_i} - \frac{Y_{..}^2}{\sum_i n_i}$ with $(t-1)$ degrees of freedom

SSW (sum of sqs. within groups) = $\sum_j \sum_i y_{ij}^2 - \sum_i \frac{(Y_{i.})^2}{n_i}$ with $(\sum_i n_i - t)$ degrees of freedom

Consequently, the total sum of squares is $TSS = SSB + SSW$. The ANOVA table can then be reformulated as follows -

Source	d.o.f.†	SS†	MS† = $\frac{SS}{d.o.f.}$	F_{cal}
between groups	$t-1$	SSB	MSB	$\frac{MSB}{MSW}$
within groups	$(\sum_i n_i - t)$	SSW	MSW	
total	$\sum_i n_i - 1$	TSS		

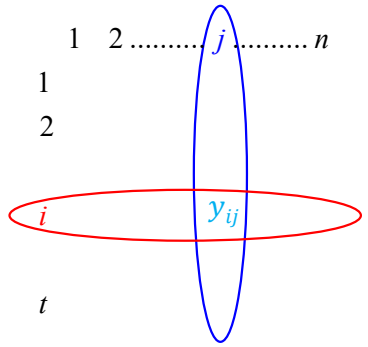
† d.o.f. means degrees of freedom, SS means sum of squares, MS means mean sum of squares.

If $F_{cal} > F_\alpha(t-1, t(n-1))$ then we reject H_0 in favour of H_1 as before

6.6.10 *Example: Rice yield across varieties*

An experiment to compare the yield of four varieties of rice is conducted. Each of the plots on a test farm where soil fertility is fairly homogeneous is treated alike relative to water and fertilizer. Four plots are randomly assigned each of the four varieties of rice. The yield in kg/acre is recorded for each plot for this randomized experiment. Does the data presented in the following table indicate a difference in the mean yield between the four varieties? Choose $\alpha = 0.01$.

variety	yield			
1	934	1041	1028	935
2	880	963	924	946
3	987	951	976	840
4	992	1143	1140	1191



Solution:

The hypothesis is stated as follows.

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1: \quad \text{not all varieties have the same mean.}$$

Here μ_i denotes the mean yield of the i^{th} variety. $n = 4$, $t = 4$. The one-way ANOVA table is printed below.

Rice variety	yield data	$Y_{i.}$ (totals)	$\overline{y_{i.}}$ (means)	SS_i
1	934 1041 1028 935	3938	984.50	10085
2	880 963 924 946	3713	928.25	3868.75
3	987 951 976 840	3754	938.50	13617
4	992 1143 1140 1191	4466	1116.5	22305
	overall	15871	991.94	49875.75

$ns^2_{means} = n\frac{\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2}{t-1} = 29977.06$. Further, $s^2_p = \frac{\sum_i SS_i}{t(n-1)} = 4156.31$.

$$F_{cal} = \frac{29977.06}{4156.31} = 7.21.$$

Using the F distribution table or using Matlab, we can find $F_{0.01}(3,12) = 5.95$. Since $F_{cal} > F_{0.01}(3,12)$, we reject H_0 and infer that there is significant difference in yield between the different rice yield.

We would arrive at the same conclusion if we had used the alternate approach to perform the calculation. In the alternate approach, the one-way ANOVA table is as follows.

Source	d.o.f.	SS	MS = $\frac{SS}{d.o.f.}$	F_{cal}
between varieties	3	89931.19	29977.06	$\frac{MSB}{MSW} = 7.21$
within varieties	12	49875.75	4156.31	
total	15	139806.94		

$$SS_1 = (934-984.5)^2 + (1041-984.5)^2 + (1028-984.5)^2 + (935-984.5)^2 = 10085$$

$$SS_2 = \ldots\ldots\ldots\ldots = 3868.75$$

$$SS_3 = \ldots\ldots\ldots\ldots = 13617$$

$$SS_4 = (992-1116.5)^2 + (1143-1116.5)^2 + (1140-1116.5)^2 + (1191-1116.5)^2 = 22305$$

$$s^2_p = \frac{10085+3868.75+13617+22305}{4(4-1)} = 4156.31$$

$$ns^2_{means} = 4\left[\frac{(984.5-991.94)^2 + \ldots.. + (1116.5-991.94)^2}{4-1}\right] = 29977.06$$

$$\sum_i \frac{(Y_{i.})^2}{n_i} - \frac{Y^2_{..}}{\sum_i n_i} = \frac{3938^2+3713^2+3754^2+4466^2}{4} - \frac{15871^2}{16}$$

$$= \frac{63331885}{4} - \frac{15871^2}{16} = 15832971 - 15743040 = 89931$$

$$\sum_j \sum_i y^2_{ij} - \sum_i \frac{Y^2_{i.}}{n_i} = 4156.31$$

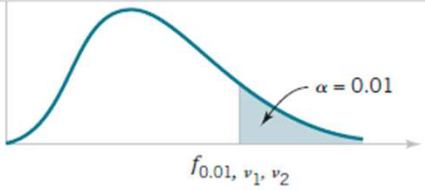


TABLE VI Percentage Points f_{α,ν_1,ν_2} of the F Distribution (*continued*)

f_{α,ν_1,ν_2}

ν_2 \ ν_1	Degrees of freedom for the numerator (ν_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06

Degrees of freedom for the denominator (ν_2)

Unequal Population Sizes (Data sets are not balanced)

Though this can be done, this unbalanced approach is generally not recommended. Whenever possible, choosing a balanced design is recommended.

The test statistic in a balanced design tends to be relatively insensitive to slight departures from the assumption of equal population variances, i.e., the balanced design tends to be more robust!