# ***Part 1:*** *Probability Distributions*

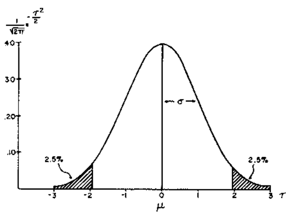
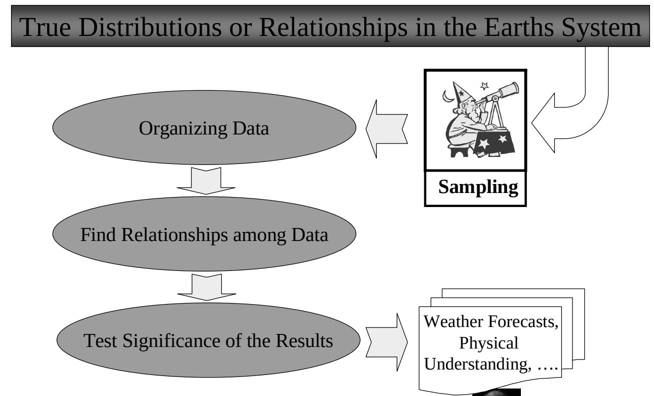


FIGURE 5
The Normal Distribution.

- ❑ Probability Distribution
- ❑ Normal Distribution
- ❑ Student-***t*** Distribution
- ❑ Chi Square Distribution
- ❑ *F* Distribution
- ❑ Significance Tests



# Purposes of Data Analysis






# Parameters and Statistics

- ❑ ***Parameters***: Numbers that describe a population. For example, the population mean (μ)and standard deviation (σ).

  ***Statistics***: Numbers that are calculated from a sample.
- ❑ A given population has only one value of a particular parameter, but a particular statistic calculated from different samples of the population has values that are generally different, both from each other, and from the parameter that the statistics is designed to estimate.
- ❑ The science of statistics is concerned with how to draw reliable, quantifiable inferences from statistics about parameters.

(from Middleton 2000)

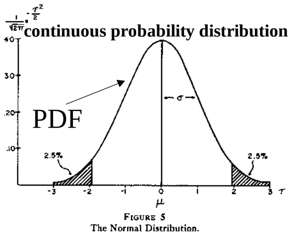


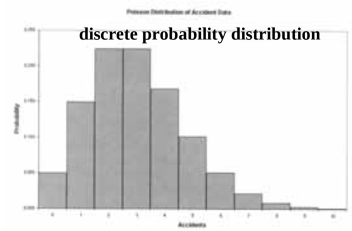
# Variables and Samples

- ❑ **Random Variable**: A variable whose values occur at random, following a probability distribution.
- ❑ **Observation**: When the random variable actually attains a value, that value is called an observation (of the variable).
- ❑ **Sample**: A collection of several observations is called sample. If the observations are generated in a random fashion with no bias, that sample is known as a random sample.

By observing the distribution of values in a random sample, we can draw conclusions about the underlying probability distribution.

# Probability Distribution




**The pattern of probabilities for a set of events is called a probability distribution.**

(1) The probability of each event or combinations of events must range from 0 to 1.

(2) The sum of the probability of all possible events must be equal too 1.



# Example

- If you throw a die, there are six possible outcomes: the numbers 1, 2, 3, 4, 5 or 6. This is an example of a random variable (the dice value), a variable whose possible values occur at random.
- When the random variable actually attains a value (such as when the dice is actually thrown) that value is called an observation.
- If you throw the die 10 times, then you have a random sample which consists of 10 observations.



# Probability Density Function

$$P(a \le x \le b) = \int_a^b f(x)\ dx$$

- ***P*** = the probability that a randomly selected value of a variable X falls between a and b.

  ***f(x)*** = the probability density function.
- The probability function has to be integrated over distinct limits to obtain a probability.
- The probability for X to have a particular value is ZERO.
- Two important properties of the probability density function:

  (1) $f(x) \ge 0$ for all x within the domain of *f*.

  (2) $\int_{-\infty}^{\infty} f(x)dx = 1$



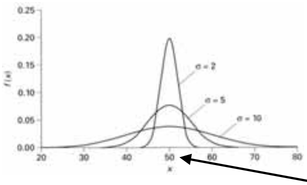
# Cumulative Distribution Function

$$F(x) \ = \ \int_{-\infty}^{x} f(t)dt$$

- The cumulative distribution function *F(x)* is defined as the probability that a variable assumes a value less than *x*.
- The cumulative distribution function is often used to assist in calculating probability (will show later).
- The following relation between F and P is essential for probability calculation:

$$P(a \le x \le b) = F(b) - F(a)$$

# Normal Distribution



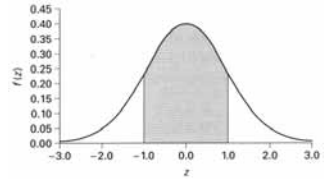

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right\}$$

**mean**

- ***f***: probability density function
  μ: mean of the population
  σ: standard deviation of the population
- The normal distribution is one of the most important distribution in geophysics. Most geophysical variables (such as wind, temperature, pressure, etc.) are distributed normally about their means.

ESS210B
Prof. Jin-Yi Yu

# Standard Normal Distribution




$$P(z_1 \le z \le z_2) = \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-z^2/2}dz$$

- The standard normal distribution has a mean of 0 and a standard deviation of 1.
- This probability distribution is particularly useful as it can represent any normal distribution, whatever its mean and standard deviation.
- Using the following transformation, a normal distribution of variable X can be converted to the standard normal distribution of variable Z:

$$\boxed{Z = (X - \mu) / \sigma}$$

ESS210B
Prof. Jin-Yi Yu

# Transformations

- It can be shown that any frequency function can be transformed in to a frequency function of given form by a suitable transformation or functional relationship.
- For example, the original data follows some complicated skewed distribution, we may want to transform this distribution into a known distribution (such as the normal distribution) whose theory and property are well known.
- Most geoscience variables are distributed normally about their mean or can be transformed in such a way that they become normally distributed.
- The normal distribution is, therefore, one of the most important distribution in geoscience data analysis.

ESS210B
Prof. Jin-Yi Yu

# How to Use Standard Normal Distribution

Table 3.3. *Normal probability table giving the area under the standard normal curve between $z=-\infty$ and $z=z_1$.*

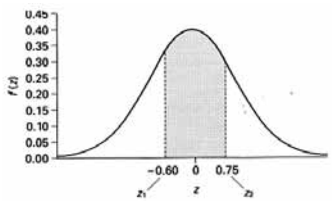
| $z_1$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0.00** | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| **0.10** | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| **0.20** | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| **0.30** | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| **0.40** | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| **0.50** | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| **0.60** | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| **0.70** | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| **0.80** | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| **0.90** | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| **1.00** | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |

- Example 1: What is the probability that a value of Z is greater than 0.75?

  Answer: P(Z≥0.75) = 1-P(Z≤0.75)=1-0.7734=0.2266

ESS210B
Prof. Jin-Yi Yu

# Another Example




***negative value***

- Example 2: What is the probability that Z lies between the limits $Z_1$=-0.60 and $Z_2$=0.75?

Answer:

P(Z ≤-0.60) = 1 – P(Z < 0.60) ➔ P(Z >-0.60)=1–0.7257=0.2743

P(-0.60 ≤Z ≤0.75) = P(Z ≤0.75) – P(Z ≤-0.60)
= 0.7734 – 0.2743
= 0.4991

ESS210B
Prof. Jin-Yi Yu

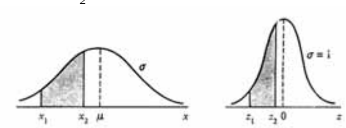
# Example 1

Given a normal distribution with μ = 50 and σ = 10, find the probability that X assumes a value between 45 and 62.

The Z values corresponding to $x_1$ = 45 and $x_2$ = 62 are

$$z_1 = \frac{45 - 50}{10} = -0.5$$

$$z_2 = \frac{62 - 50}{10} = 1.2.$$



Therefore, P(45 < *X* < 62)
= P(-0.5 < *Z* < 1.2)
= P(Z <1.2) – P(Z<-0.5)
= 0.8849 – 0.3085
= 0.5764

ESS210B
Prof. Jin-Yi Yu

# Example 2

An electrical firm manufactures light bulbs that have a length of life that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the possibility that a bulb burns between 778 and 834 hours.

The Z values corresponding to $x_1$ = 778 and $x_2$ = 834 are

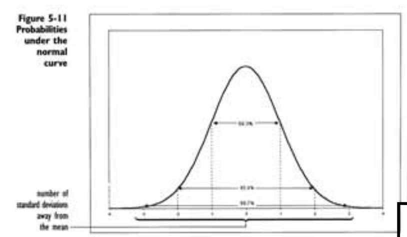
$z_1$ = (778 – 800) / 40 = -0.55

$z_2$ = (834 – 800) / 40 =0.85

Therefore, P(778 < *X* < 834)
= P(-0.55 < *Z* < 0.85)
= P(Z <0.85) – P(Z<-0.55)
= 0.8849 – 0.2912
= 0.5111

ESS210B
Prof. Jin-Yi Yu

# Probability of Normal Distribution




$$P(-1 \le z \le 1) = \int_{-1}^{+1} f(z)\,dz = 68.27\%$$

$$P(-2 \le z \le 2) = \int_{-2}^{+2} f(z)\,dz = 95.45\%$$

$$P(-3 \le z \le 3) = \int_{-3}^{+3} f(z)\,dz = 99.73\%$$

## Probability of Normal Distribution

$$P(-1 \le z \le 1) = \int_{-1}^{+1} f(z)\,dz = 68.27\%$$

$$P(-2 \le z \le 2) = \int_{-2}^{+2} f(z)\,dz = 95.45\%$$

$$P(-3 \le z \le 3) = \int_{-3}^{+3} f(z)\,dz = 99.73\%$$

- There is only a 4.55% probability that a normally distributed variable will fall more than 2 standard deviations away from its mean.
- This is the two-tailed probability. The probability that a normal variable will exceed its mean by more then 2 $\sigma$ is only half of that, 2.275%.

ESS210B
Prof. Jin-Yi Yu

## Application in Operational Climatology

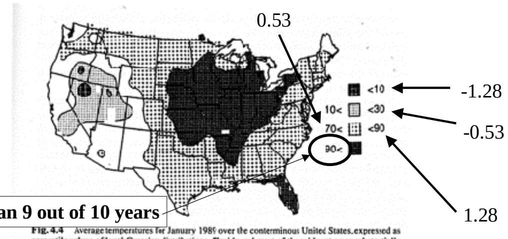



Fig. 4.4 Average temperatures for January 1989 over the conterminous United States, expressed as percentile values of local Gaussian distributions. Florida and most of the midwest were substantially warmer than usual, while only a small portion of the western United States was cooler than usual. [From Arkin (1989), reproduced with permission of the American Meteorological Society.]

- This figure shows average temperatures for January 1989 over the US, expressed as quantiles of the local Normal distributions.
- Different values of $\mu$ and $\sigma$ have been estimated for each location.

ESS210B
Prof. Jin-Yi Yu

## How to Estimate $\mu$ and $\sigma$

- In order to use the normal distribution, we need to know the mean and standard deviation of the population

→ But they are impossible to know in most geoscience applications, because most geoscience populations are infinite.

→ We have to estimate $\mu$ and $\sigma$ from samples.

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{and} \quad s = \left[\frac{\sum (x_i - \bar{x})^2}{n-1}\right]^{\frac{1}{2}}$$

ESS210B
Prof. Jin-Yi Yu

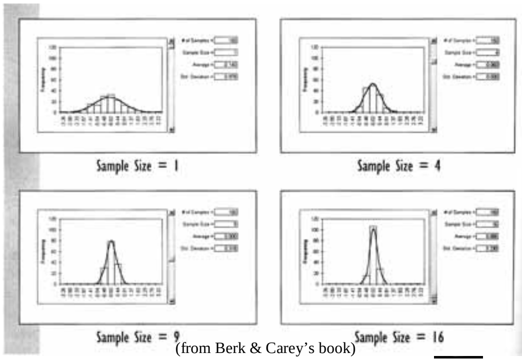
## Sampling Distribution

- Is the sample mean close to the population mean?
- To answer this question, we need to know the probability distribution of the sample mean.
- We can obtain the distribution by repeatedly drawing samples from a population and find out the frequency distribution.

ESS210B
Prof. Jin-Yi Yu

# Example

- ❑ If we repeatedly draw N observations from a standard normal distribution ($\mu$=0 and $\sigma$=1) and calculate the sample mean, what is the distribution of the sample mean?




(from Berk & Carey's book)



# Distribution of Sample Average

❑ What did we learn from this example?

➔If a sample is composed of N random observations from a normal distribution with ***mean*** $\mu$ and ***standard deviation*** $\sigma$, then the distribution of the sample average will be a normal distribution with ***mean*** $\mu$ but ***standard deviation*** $\sigma$***/sqrt(N).***



# Standard Error

❑ The standard deviation of the probability distribution of the sample mean ($\overline{X}$) is also referred as the "standard error" of $\overline{X}$.



# Distribution of Sample Means

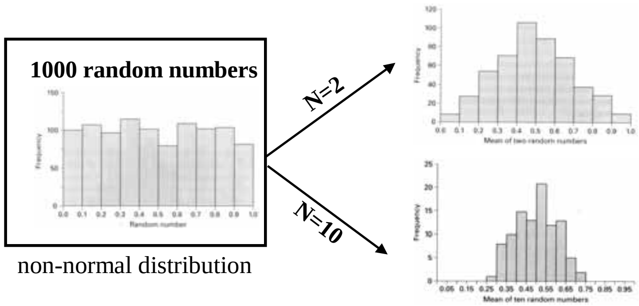
❑ One example of how the normal distribution can be used for data that are "non-normally" distributed is to determine the distribution of sample means.

❑ Is the sample mean close to the population mean?

❑ To answer this question, we need to know the probability distribution of the sample mean.

❑ We can obtain the distribution by repeatedly drawing samples from a population and find out the frequency distribution.

# An Example of Normal Distribution



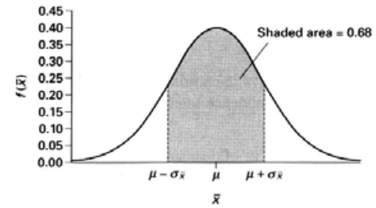
non-normal distribution

- ❑ If we repeatedly draw a sample of N values from a population and calculate the mean of that sample, the distribution of sample means tends to be in normal distribution.

**(Figure from Panofsky and Brier 1968)**



# Central Limit Theory



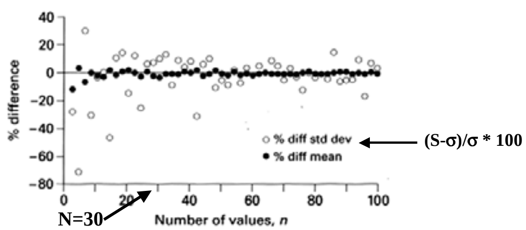

- ❑ In the limit, as the sample size becomes large, the sum (or the mean) of a set of independent measurements will have a normal distribution, irrespective of the distribution of the raw data.
- ❑ If there is a population which has a mean of μand a standard deviation of σ, then the distribution of its sampling means will have:

(1) a mean $\mu_{\bar{x}} = \mu$, and

(2) a standard deviation $\sigma_{\bar{x}} = \sigma/(N)^{1/2}$ = standard error of the mean



# Sample Size




- ❑ The bigger the sample size N, the more accurate the estimated mean and standard deviation are to the true values.
- ❑ How big should the sample size is for the estimate mean and standard deviation to be reasonable?
  - ➔ N > 30 for *s* to approach σ (μ gets approached even with small N)



# Normal Distribution of Sample Means

- ❑ The distribution of sample means can be transformed into the standard normal distribution with the following transformation:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \left(\frac{\bar{x} - \mu}{s/\sqrt{n}}\right)$$

- ❑ The confidence level of the sample mean can therefore be estimated based on the standard normal distribution and the transformation.

## The Importance of the Central Limit Theorem

- With this theorem, statistician can make reasonable inferences about the sample mean without having to know the underlying probability distribution.

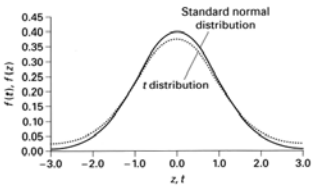


## Small Sampling Theory

- When the sample size is smaller than about 30 (N<30), we can not use the normal distribution to describe the variability in the sample means.
- With small sample sizes, we cannot assume that the sample-estimated standard deviation (***s***) is a good approximation to the true standard deviation of the population ($\sigma$).
- The quantity $\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right)$ no longer follows the standard normal distribution.
- In stead, the probability distribution of the small sample means follows the "Student's ***t*** distribution".



## Student's *t* Distribution




- As the sample size (n) increases, the t-distribution approaches the normal distribution.
- Unlike the normal distribution, the ***t***-distribution depends on sample size. Its tails get longer with smaller n.
- For variable ***t***

$$t=\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right)$$

- The probability density function for the ***t***-distribution is:

$$f(t)=K(\nu)\left(1+\frac{t^2}{n-1}\right)^{-n/2}$$

- Here K($\nu$) is a constant which depends on the number of degree of freedom ($\nu$ = n-1). K($\nu$) is chosen so that:

$$\int_{-\infty}^{\infty} f(t)dt=1$$



## How to Use *t*-Distribution

Table 3.9. *t values for various confidence levels and degrees of freedom.*[30]

| Number of values, $n$ | Degrees of freedom, $\nu$ | $t_{68\%,\nu}$ | $t_{90\%,\nu}$ | $t_{95\%,\nu}$ | $t_{99\%,\nu}$ |
|---|---|---|---|---|---|
| 2 | 1 | 1.82 | 6.31 | 12.71 | 63.66 |
| 4 | 3 | 1.19 | 2.35 | 3.18 | 5.84 |
| 10 | 9 | 1.05 | 1.83 | 2.26 | 3.25 |
| 20 | 19 | 1.02 | 1.73 | 2.09 | 2.86 |
| 30 | 29 | 1.01 | 1.70 | 2.05 | 2.76 |
| 50 | 49 | 1.00 | 1.68 | 2.01 | 2.68 |
| 100 | 99 | 1.00 | 1.66 | 1.98 | 2.63 |
| 10000 | 9999 | 0.99 | 1.65 | 1.96 | 2.58 |

# An Example

In a sample of 10 winters the mean January temperature is 42˚F and the standard deviation is 5˚F. What are the 95% confidence limits on the true mean January temperature?

1. Desired confidence level is 95%.

2. The null hypothesis is that the true mean is between $42 \pm \Delta T$. The alternative is that it is outside this region.

3. We will use the $t$ statistic.

Sample size is small

4. The critical region is $|t| < t_{.025}$, which for $n = N - 1 = 9$ is $|t| < 2.26$. Stated in terms of confidence limits on the mean we have:

$$\bar{x} - 2.26 \cdot \frac{s}{\sqrt{N-1}} < \mu < \bar{x} + 2.26 \cdot \frac{s}{\sqrt{N-1}}$$

5. Putting in the numbers we get $38.23 < \mu < 45.77$. We have 95% certainty that the true mean lies between these values. This is the answer we wanted. If we had a guess about what the true mean was, we could say whether the data would allow us to reject this null hypothesis at the significance level stated.

(From Hartmann 2003)



# Student's *t* and Normal Distributions

- Since the Student's *t* distribution approaches the normal distribution for large *N*, there is no reason to use the normal distribution in preference to Student's *t*.

- The Student 's *t* distribution is the most commonly used in meteorology, and perhaps in all of applied statistics.



# Degree of Freedom

- If you are asked to choose a pair of number ($x_1$, $x_2$) at random, you have complete freedom of choice with regard to each of the two numbers.
  - ➔ There are two degrees of freedom.
- If somehow you are asked to choose a pair of numbers whose sum is 7 ($x_1+x_2=7$) , you are free to choose only one number (for example, $x_1$). The other number has to be $x_2=7-x_1$.
  - ➔ There is only one degree of freedom.
- If you are further asked to make sure $x_1^2+x_2^2=5$, then you have no freedom to choose the number. $X_1$ has to be 3 and $x_2$ has to be 4.
  - ➔ The degree of freedom is zero.



# How to Determine DF?

- The number of degree of freedom
  = (the number of original observations) – (the number of parameters estimated from the observation)

- For example, the number of degree of freedom for a variance is N-1, where N is the number of the observations.

$$Variance = \Sigma(X - \bar{X})^2$$

- It is because we need to estimate the mean (a parameter) from the observations, in order to calculate the variance.

# Example

- We have four observations: [4, 1, 8, 3]
- To calculate the variance, we first get the mean = (4+1+8+3)/4=4
- We then calculate the deviations of each observations from the mean and get d = [4-4, 1-4, 8-4, 3-4] = [0, -3, 4, -1].
- Now these four d's are constrained by the requirement:
  d1 + d2 + d3 + d4 =0 (a result related to the mean).
- The variance we get (variance = d1**2+d2**2+d3**2+d4**2) has only 3 degree of freedom ( 4 observations minus one estimate of the parameter).



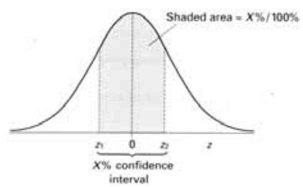
# Statistical Inference

Two of the main tools of statistical inference are:

- Confidence Intervals
  Within what intervals (or limits) does X% of the population lie in the distribution
- Hypothesis Tests
  You formulate a theory (hypothesis) about the phenomenon you are studying and examine whether the theory is supported by the statistical evidence.



# Confidence Intervals (or Limits)




- In stead of asking "what is the probability that Z falls within limits *a* and *b* in the normal distribution", it is more important to ask "Within what intervals or limits does X% of the population lie in the normal distribution".
- The X% is referred as the "confidence level".
- The interval correspond to this level is called the "confidence intervals" or "confidence limits".



# Normal Distribution of Sample Means

- The distribution of sample means can be transformed into the standard normal distribution with the following transformation:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \left(\frac{\bar{x} - \mu}{s/\sqrt{n}}\right)$$

Z-test statistics

- The confidence level of the sample mean can therefore be estimated based on the standard normal distribution and the transformation.

## How to Determine Confidence Intervals

**Table 3.5.** *z values corresponding to X% confidence levels.*

| $X\%$ | 50% | 68% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $z_{X\%}$ | 0.67 | 0.99 | 1.64 | 1.96 | 2.58 |

- Choose a confidence level
- Determine ***Z-values*** from the table of standard normal distribution. (Note that $Z_1 = - Z_2$)
- Transform Z values back to X values using $Z = (X - \mu) / \sigma$
- The Confidence intervals are determined.

ESS210B
Prof. Jin-Yi Yu

## Confidence Intervals for Sample Mean

- The probability that a sample mean, $\underline{x}$, lies between $\mu$-$\sigma_{\underline{x}}$ and $\mu$+ $\sigma_{\underline{x}}$ may be written:

$$P\left(\mu-\frac{\sigma}{\sqrt{n}}\le \bar{x}\le \mu+\frac{\sigma}{\sqrt{n}}\right)=0.68$$

- However, it is more important to find out the confidence interval for the population mean $\mu$.. The previous relation can be written as:

$$P\left(\bar{x}-\frac{\sigma}{\sqrt{n}}\le \mu\le \bar{x}+\frac{\sigma}{\sqrt{n}}\right)=0.68$$

- But we don't know the true value of $\sigma$ (the standard deviation of the population). *If the sample size is large enough ($N \ge 30$)*, we can use ***s*** (the standard deviation estimated by samples) to approximate $\sigma$:

$$P\left(\bar{x}-\frac{s}{\sqrt{n}}\le \mu\le \bar{x}+\frac{s}{\sqrt{n}}\right)=0.68$$

ESS210B
Prof. Jin-Yi Yu

## Confidence Intervals for Sample Mean

- The 68% confidence intervals for $\mu$ are:

$$\bar{x}-\frac{s}{\sqrt{n}}\le \mu\le \bar{x}+\frac{s}{\sqrt{n}}$$

- To generalize the relation to any confidence level X%, the previous relation can be rewritten as:

$$\bar{x}-z_{X\%}\frac{s}{\sqrt{n}}\le \mu\le \bar{x}+z_{X\%}\frac{s}{\sqrt{n}}$$

**Table 3.5.** *z values corresponding to X% confidence levels.*

| $X\%$ | 50% | 68% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| $z_{X\%}$ | 0.67 | 0.99 | 1.64 | 1.96 | 2.58 |

ESS210B
Prof. Jin-Yi Yu

## An Example

- In an experiment, forty measurements of air temperature were made. The mean and standard deviation of the sample are:

$\underline{x}$ = 18.41°C and ***s*** = 0.6283°C.

***Question 1:***

Calculate the 95% confidence interval for the population mean for these data.

***Answer:***

From the previous slid, we know the interval is: $\underline{x} \pm Z_{95\%} * s / (N)^{0.5}$

$Z_{95\%} = 1.96$

$Z_{95\%} * s / (N)^{0.5} = 1.96 * 0.6283°C / (40)^{0.5} = 0.1947°C$

The 95% confidence level for the population mean: 18.22°C ~ 18.60°C.

ESS210B
Prof. Jin-Yi Yu

## An Example – cont.

- In an experiment, forty measurements of air temperature were made. The mean and standard deviation of the sample are:

  $\underline{x}$ = 18.41°C and $s$ = 0.6283°C.

***Question 2:***

How many measurements would be required to reduce the 95% confidence interval for the population mean to ($\underline{x}$-0.1)°C to ($\underline{x}$+0.1)°C?

***Answer:***

We want to have $Z_{95\%} * s / (N)^{0.5} = 0.1$ °C

We already know $Z_{95\%} = 1.96$ and $s$ = 0.6283°C

➔ $N = (1.96 \times 0.6283/0.1)^2 = 152$

ESS210B
Prof. Jin-Yi Yu

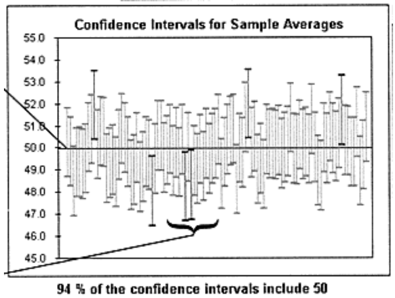
## Application of Confidence Intervals

- A typical use of confidence intervals is to construct error bars around plotted sample statistics in a graphical display.

ESS210B
Prof. Jin-Yi Yu

## Interpreting the Confidence Intervals

- The term 95% confident means that we are confident our procedure will capture the value of μ 95% of the times it (the procedure $\bar{x} - \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}}$ ) is used.




95% confidence intervals calculated from 100 samples, each of which contains 25 observations.

(from Bock & Lapey's book)

ESS210B
Prof. Jin-Yi Yu

## Hypothesis (Significance) Testing

| Population parameter | Sample statistic |
|---|---|
| Mean, $\mu$ | Mean, $\bar{x}$ |
| Standard deviation, $\sigma$ | Standard deviation, $s$ |
| Correlation coefficient, $\rho$ | Correlation coefficient, $r$ |
| Intercept, $\alpha$ (of a line through $x$–$y$ data) | Intercept, $a$ |
| Slope, $\beta$ (of a line through $x$–$y$ data) | Slope, $b$ |

- Hypothesis testing involves comparing a hypothesized population parameter with the corresponding number (or statistic) determined from sample data.
- The hypothesis testing is used to construct confidence interval around sample statistics.
- The above table lists the population parameters that are often tested between the population and samples.

ESS210B
Prof. Jin-Yi Yu

## Parametric .vs. Nonparametric Tests

- **Parametric Tests**: conducted in the situations where one knows or assumes that a particular theoretical distribution (e.g., Normal distribution) is an appropriate representation for the data and/or the test statistics.

→ In these tests, inferences are made about the particular distribution parameters.

- **Nonparametric Test**: conducted without the necessity of assumption about what theoretical distribution pertains to the data.



## Significance Tests

- Five Basic Steps:
  1. State the null hypothesis and its alternative
  2. State the statistics used
  3. State the significance level
  4. State the critical region
     (i.e., identify the sample distribution of the test statistic)
  5. Evaluate the statistics and state the conclusion



## Null Hypothesis

- Usually, the null hypothesis and its alternative are mutually exclusive. For example:

$H_0$: The means of two samples are equal.
$H_1$: The means of two samples are not equal.

$H_0$: The variance at a period of 5 days is less than or equal to C.
$H_1$: The variance at a period of 5 days is greater than C.

Hypotheses that we formulate with the hope of rejecting are called null hypothesis and are denoted by ***Ho***.



## Examples

- If a researcher is testing a new cold vaccine, he/she should assume that it is no better than the vaccine now on the market and then set out to reject this contention.
- To prove that one technology is better than another, we test the hypothesis that there is no difference in these two techniques.

## Rejection/Acceptance of Hypothesis

- Evidences from the sample that is inconsistent with the stated hypothesis leads to the rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance.
- The rejection of a hypothesis is to conclude it is false.
- The acceptance of a statistical hypothesis is a result of insufficient evidence to reject it and does not necessary imply that it is true.
- Therefore, it is better to state a hypothesis that we wish to reject.

Hypotheses that we formulate with the hope of rejecting are called null hypothesis and are denoted by ***Ho***.



## Test Statistics

- A test statistic is the quantity computed from the sample that will be the subject of the test.
- In parametric tests, the test statistic will often be the sample estimate of a parameter of a relevant theoretical distribution.



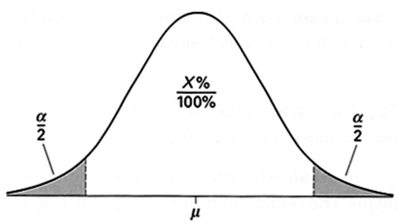
## Example

- To examine the significance of the difference between means obtained from two different samples. We use the "two-sample t-statistic":

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$



## Confidence and Significance




- Confidence Level: X%
- Significance Level: $\alpha$

## What Does It Mean 5% Significant?

- At the 5% significance level, there is a one in twenty chance of rejecting the hypothesis wrongly (i.e., you reject the hypothesis but it is true).

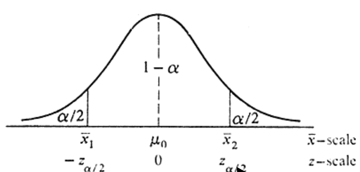


## Sampling Distribution

- A statistic is calculated from a batch of data.
- The value of the statistic varies from batch to batch.
- A sampling distribution for a statistic is the probability distribution describing batch-to-batch variations of that statistic.
- The random variations of sample statistics can be described using probability distributions just as the random variations of the underlying data can be described using probability distributions.
- Sample statistics can be viewed as having been drawn from probability distributions.



## Critical Regions




The location of the critical region can be determined only after $H_1$ has been stated.

| $H_0$ | Statistics | $H_1$ | Critical Region |
|---|---|---|---|
| $\mu = \mu_0$ | $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$; $\sigma$ known | $\mu < \mu_0$ | $Z < -z_\alpha$ |
| | | $\mu > \mu_0$ | $Z > z_\alpha$ |
| | | $\mu \neq \mu_0$ | $Z < -z_{\alpha/2}$ and $Z > z_{\alpha/2}$ |



## p Value

- The p value is the specific probability that the observed value of the test statistic will occur according to the null distribution (the probability distribution based on the null hypothesis).
- If p value falls within the critical regions
  - ➔ Reject the null hypothesis.
- The p value depends on the alternative hypothesis.

## Example

- Ho: μ = 50
  H1: μ ≠ 50

If a N=25 sampling mean has a sample mean of 45 and a sample σ of 15, then

Z = (45-50) / (15/5) = -1.67
➔ P(Z<-1.67)=0.0478
➔ p value = 2 * 0.0478 = 0.0956

- Ho: μ = 50
  H1: μ < 50

Z = (45-50) / (15/5) = -1.67
➔ P(Z<-1.67)=0.0478
➔ p value = 0.0478 = 0.0478

ESS210B
Prof. Jin-Yi Yu

## Rejection/Acceptance of Hypothesis

- Evidences from the sample that is inconsistent with the stated hypothesis leads to the rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance.
- The rejection of a hypothesis is to conclude it is false.
- The acceptance of a statistical hypothesis is a result of insufficient evidence to reject it and does not necessary imply that it is true.
- Therefore, it is better to state a hypothesis that we wish to reject.

Hypotheses that we formulate with the hope of rejecting are called null hypothesis and are denoted by ***Ho***.

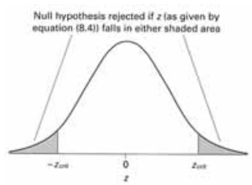
ESS210B
Prof. Jin-Yi Yu

## Type-I and Type-II Errors

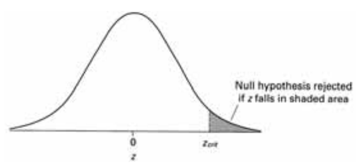
- Type I Error: You reject a null hypothesis that is true.
- Type II Error: You fail to reject a null hypothesis that is false.
- By choosing a significance level of, say, α=0.05, we limit the probability of a type I error to 0.05.

ESS210B
Prof. Jin-Yi Yu

## One-Side and Two-Side Tests



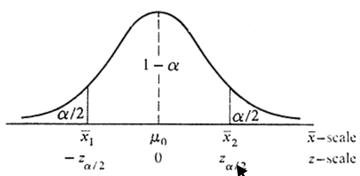

- Two-Side Tests
  $H_0$: mean temperature = 20°C
  $H_1$: mean temperature ≠ 20°C




- One-Side Tests
  $H_0$: mean temperature = 20°C
  $H_1$: mean temperature > 20°C

ESS210B
Prof. Jin-Yi Yu

# Critical Regions




The location of the critical region can be determined only after $H_1$ has been stated.

| $H_0$ | Statistics | $H_1$ | Critical Region |
|---|---|---|---|
| $\mu = \mu_0$ | $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$; $\sigma$ known | $\mu < \mu_0$ | $Z < -z_\alpha$ |
| | | $\mu > \mu_0$ | $Z > z_\alpha$ |
| | | $\mu \neq \mu_0$ | $Z < -z_{\alpha/2}$ and $Z > z_{\alpha/2}$ |

ESS210B
Prof. Jin-Yi Yu

# Example

- A manufacturer of sports equipment has developed a new fishing line that he claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kilograms against the alternative that $\mu \neq 8$ kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilogram. Use a 0.01 level of significance.

***Answer:***

1. $\alpha = 0.01$
2. $H_0$: $\mu = 8$ kilograms
   $H_1$: $\mu \neq 8$ kilograms
3. Use Z statistics (because sample mean has a normal distribution)
   $$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{N}}}$$
4. Critical region: $Z_{-\alpha/2} < -2.58$ and $Z_{\alpha/2} > 2.58$
5. $Z = (7.8-8)/0.5/(50)^{0.5} = -2.828 < Z_{-\alpha/2}$
6. Reject $H_0$ and conclude that the average breaking strength is not equal to 8 but, in fact, less than 8 kilograms.

ESS210B
Prof. Jin-Yi Yu

# Significance Test of the Difference of Means

- We want to compare the averages from two independent samples to determine whether a significance exists between the samples.
- For Example:
  * One sample contains the cholesterol data on patients taking a standard drug, while the second sample contains cholesterol data on patients taking experimental drug. You would test to see whether there is statistically significant difference between two sample averages.

  * Compare the average July temperature at a location produced in a climate model under a doubling versus no doubling CO2 concentration.

  * Compare the average winter 500-mb height when one or the other of two synoptic weather regimes had prevailed.

ESS210B
Prof. Jin-Yi Yu

# Two kinds of Two-Sample t-Test

- For samples come from distributions with different standard deviations, having values of $\sigma_1$ and $\sigma_2$. This is called "unpooled two-sample t-statistic":

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- If both distributions have the same standard deviation, then we can "pool" the estimates of the standard deviation from the two samples into a single estimate:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\sigma} \qquad \sigma = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

ESS210B
Prof. Jin-Yi Yu

# Paired Data (One-Sample) t-Test

- Paired Data: where observations come in natural pairs.
  For example:
  A doctor might measure the effect of a drug by measuring the physiological state of patients before and after applying the drug. Each patient in this study has two observations, and the observations are paired with each other. To determine the drug's effectiveness, the doctor looks at the difference between the before and the after reading.

- Paired-data test is a one-sample t-test
  To test whether or not there is a significant difference, we use one-sample t-test. It is because we are essentially looking at one sample of data – the sample of paired difference.



# Example

- There is a data set that contains the percentage of women in the work force in 1968 and 1972 from a sample of 19 cities. There are two observations from each city, and the observations constitute paired data. You are asked to determine whether this sample demonstrates a statistically significant increase in the percentage of women in the work force.
- **Answer:**

1. $H_o$: $\mu = 0$ (There is no change in the percentage)
   $H_1$: $\mu \neq 0$ (There is some change, but we are not assuming the direction of change)
2. 95% t-Test with N=19 ➔ We can determine $t_{0.25\%}$ and $t_{0.975\%}$
3. S=0.05974 ➔ $\sigma = S/(19)^{0.5} = 0.01371$
4. t statistic t = (**0.0337** – 0) / $\sigma$ = 2.458 > $t_{0.975\%}$

**mean increase in the Percentage from 19 cities**

5. We reject the null hypothesis ➔ There has been a significant change in women's participation in the work force in those 4 years.

ESS210B
Prof. Jin-Yi Yu

| City | Year 1968 | Year 1972 | Difference |
|---|---|---|---|
| New York | 0.42 | 0.45 | 0.03 |
| Los Angeles | 0.50 | 0.50 | 0.00 |
| Chicago | 0.52 | 0.52 | 0.00 |
| Philadelphia | 0.45 | 0.45 | 0.00 |
| ….. | …… | ……. | …… |
| ….. | …… | ……. | ……. |
| …. | ….. | …… | …… |
| ….. | ….. | ….. | ….. |
| Dallas | 0.63 | 0.64 | 0.01 |

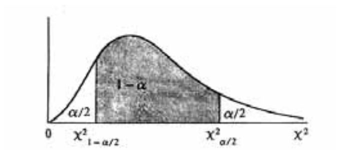
ESS210B
Prof. Jin-Yi Yu

# Test of Sampling Variance

- Standard deviation is another important population parameter to be tested.
- The sample variances have a Chi-Squared ($\chi^2$) distribution, with the following definition:

$$\boxed{\chi^2 = (N-1)\frac{s^2}{\sigma^2}}$$

statistic




- Here N: sample size
  S: the sampled standard deviation
  $\sigma$: the true standard deviation

ESS210B
Prof. Jin-Yi Yu

# $\chi^2$ Distribution

**ν=N-1**

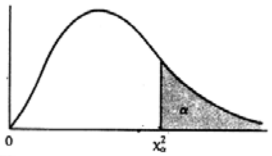



**Table VI†** Critical Values of the Chi-Square Distribution

| | α | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ν | 0.995 | 0.99 | 0.975 | 0.95 | 0.05 | 0.025 | 0.01 | 0.005 |
| 1 | $0.0^4393$ | $0.0^3157$ | $0.0^3982$ | $0.0^2393$ | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.0100 | 0.0201 | 0.0506 | 0.103 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.0717 | 0.115 | 0.216 | 0.352 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 11.070 | 12.832 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 18.307 | 20.483 | 23.209 | 25.188 |



# Example

A manufacturer of car batteries claims that the life of his batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

1. Ho: $\sigma^2 = 0.81$
   $H_1$: $\sigma^2 > 0.81$
2. $\alpha = 0.05$
3. Critical region: $\chi^2 > 16.919$, where $\chi^2 = (n-1)S^2/\sigma^2$ with degree of freedom=9
4. From the sampling: $S^2=1.44$, n=10,
   $\chi^2 = (10-1)(1.44)/0.81 = 16.0 < 16.919$
5. Conclusion: Accept Ho and conclude that these is no reason to doubt that the standard deviation is 0.9 year.



# Significance Test of the Difference of Variances

- We want to know if the standard deviation estimated from two samples are significantly different.
- The statistical test used to compare the variability of two samples is the ***F***-test:

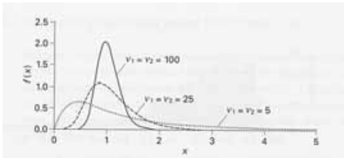
$$F = \frac{s_1^2}{s_2^2}$$

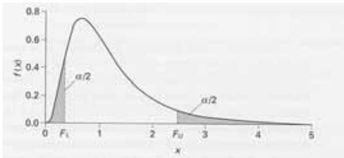
$S_1^2$: the estimate of the population variance of sample 1

$S_2^2$: the estimate of the population variance of sample 2.



# *F*-Distribution

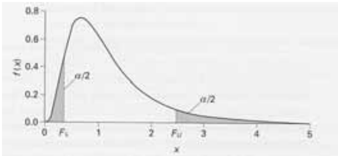



$$f(x) = K(\nu_1, \nu_2)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{\frac{\nu_1 - 2}{2}}}{\left(1 + \frac{\nu_1 x}{\nu_2}\right)^{\frac{\nu_1 + \nu_2}{2}}}$$

- Here:
  ***X* = $S_1^2/S_2^2$**
  $\nu_1$ = degree of freedom of sample 1
  $\nu_2$ = degree of freedom of sample 2
  K($\nu_1$, $\nu_2$) = a constant to make the total area under the f(x) curve to be 1

# $F$- -Distribution




- F-distribution is non-symmetric.
- There are two critical F values of unequal magnitude.
- Whether to use the upper or the lower critical value depends on the relative magnitudes of $S_1$ and $S_2$:

  If $S_1 > S_2$, then use the upper limit $F_U$ to reject the null hypothesis

  If $S_1 < S_2$, then use the lower limit $F_L$ to reject the null hypothesis
- Or you can always make $S_1 > S_2$ when F is defined. In this case, you can always use the upper limit to examine the hypothesis.

ESS210B
Prof. Jin-Yi Yu

Numerator $\nu_1$

| Denominator degrees of freedom, $\nu_2$ | $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | 0.1 | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 |
| | 0.05 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 |
| | 0.025 | 647.79 | 799.48 | 864.15 | 899.60 | 921.83 |
| 2 | 0.1 | 8.53 | 9.00 | 9.16 | 9.24 | 9.29 |
| | 0.05 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 |
| | 0.025 | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 |
| | 0.01 | 98.50 | 99.00 | 99.16 | 99.25 | 99.30 |
| | 0.005 | 198.50 | 199.01 | 199.16 | 199.24 | 199.30 |
| 3 | 0.1 | 5.54 | 5.46 | 5.39 | 5.34 | 5.31 |
| | 0.05 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 |
| | 0.025 | 17.44 | 16.04 | 15.44 | 15.10 | 14.88 |
| | 0.01 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 |
| | 0.005 | 55.55 | 49.80 | 47.47 | 46.20 | 45.39 |
| 4 | 0.1 | 4.54 | 4.32 | 4.19 | 4.11 | 4.05 |
| | 0.05 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 |
| | 0.025 | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 |
| | 0.01 | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 |
| | 0.005 | 31.33 | 26.28 | 24.26 | 23.15 | 22.46 |
| 5 | 0.1 | 4.06 | 3.78 | 3.62 | 3.52 | 3.45 |
| | 0.05 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 |
| | 0.025 | 10.01 | 8.43 | 7.76 | [illegible] | 7.15 |
| | 0.01 | 16.26 | 13.27 | 12.06 | 11.39 | [illegible] |
| | 0.005 | 22.78 | 18.31 | 16.53 | 15.56 | 14.94 |

**Significance level**

**Table of $F$-Distribution**

ESS210B
Prof. Jin-Yi Yu

# Example

- There are two samples, each has six measurements of wind speed. The first sample measures a wind speed variance of 10 (m/sec)². The second sample measures a 6 (m/sec)² variance. Are these two variance significantly different?

***Answer:*** (1) Selected a 95% significance level ($\alpha = 0.05$)

(2) $H_0$: $\sigma_1 = \sigma_2$

$H_1$: $\sigma_1 \neq \sigma_2$

(3) Use $F$-test

$F = 10/6 = 1.67$

(4) This is a two-tailed test. So choose the 0.025 significance areas.

For $\nu_1 = \nu_1 = 6\text{-}1=5$, $F_{0.975}=7.15$.

(5) Since F< $F_{0.975}$, the null hypothesis can not be rejected.

(6) Variance from these two samples are not significantly different.

ESS210B
Prof. Jin-Yi Yu