Lecture #8
Sampling Distribution

BMIR Lecture Series on Probability and Statistics

Ching-Han Hsu, Ph.D.
Department of Biomedical Engineering
and Environmental Sciences
National Tsing Hua University

Populations

Definition

A **population** consists of the totality of the observations with which we are concerned.

- The totality of observations, whether their number be finite or infinite, constitutes what we call a **population**.
- The word *population* previously referred to observations obtained from statistical studies about people.
- Today, statisticians use the term to refer to observations relevant to anything of interest, whether it be groups of people, animals, or all possible outcomes from some complicated biological or engineering system.

Populations: Examples

- If there 600 students in the school whom we classified according to blood type, we say that we have a population of size 600.
- The number of the cards in a deck, the heights of residents in a city, and the lengths of cars in a parking lot are examples of populations with finite number. The total number of observations is also a finite number.
- The observations obtained by measuring the atmospheric pressure every day or all measurements of the depth of a lake are examples of populations whose sizes are infinite.

An Observation

- Each observation in a population is a value of a random variable X having some probability distribution $f(x)$.
- For example, if one is inspecting items coming off an assembly line for detect, then each observation in the population might be a value 0 or 1 of the Bernoulli random variable X with probability distribution

  $$b(x:1,p) = p^x q^{1-x},\ x = 0,1$$

  where 0 indicates a non-defective item and 1 indicates a defective one. p is the probability of any item being defective and $q = 1 - p$.
- When we refer to the population $f(x)$, i.e, binomial or normal distributions, we mean a population whose observations are values of a random variable having the probability distribution $f(x)$.

Sampling

Definition

A **sample** is a subset of population.

- In the statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population.
- We must depend on a subset of observations from the population to help us make inferences concerning that same population.
- If our inferences are to be valid, we must obtain samples that are representative of the population.
- Any sampling procedure that produces inferences that consistently over-estimate or consistently under-estimate some characteristic of the population is said to be **biased**.

Random Sample

Definition

Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution function $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

- In a random sample, the observations are made independently and at random.
- The random variable $X_i, i = 1, \ldots, n$ represents the ith measurement or sample value that we observe.
- And $x_i, i = 1, \ldots, n$ represents the real value that we measure.

Statistics

Definition

Any function of the random variables constituting a random sample is called a **statistic**.

Statistical Inferences

- We want some methods to make decisions or to draw conclusions about a **population**.
- We need **samples from population** and utilize the information within.
- The methods can be divided into two major areas: **parameter estimation** and **hypothesis testing**.

What is statistics?

- Statistics is a function of observations or random samples.
- Statistics itself is also a random variable.
- The probability distribution of a statistics is called a **sampling distribution**.

Location Measures of a Sample

Let $X_1, X_2, \ldots, X_n$ represent n random variables.

- **Sample Mean**:

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

  Note that the statistic $\bar{X}$ assume the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- **Sample Median**:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even} \end{cases}$$

  where the observations, $x_1, x_2, \ldots, x_n$, are arranged in increasing order.

- The **sample mode** is the value of the sample that occurs most often.

Variability Measures of a Sample: Example

The variability in a sample displays how the observations spread out from the average. For example,

- Consider the following measurements, in liters, for two samples of orange juice bottled by company A and B:

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

- The sample mean and std of samples A and B:

Company	Mean	STD
A	1.0	0.047
B	1.0	0.107

- The **variability**, or the **dispersion** of the observations from the average is less for sample A than for sample B.

Variability Measures of a Sample

Let $X_1, X_2, \ldots, X_n$ represent n random variables.

- **Sample Variance**:

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right] \end{aligned}$$

- **Sample Standard Deviation**:

$$S=\sqrt{S^2}$$

- **Sample Range**: Let X_{max} denote the largest of the X_i values and X_{min} the smallest:

$$R = X_{max} - X_{min}$$

Variability Measures of a Sample

Example

Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen.

Solution

We find that $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$, *and* $n = 6$. *Hence*

$$s^2 = \frac{1}{(6)(5)}\left[(6)(171) - (31)^2\right] = \frac{13}{6}$$

The sample standard deviation $s = \sqrt{\frac{13}{6}} = 1.47$ *and the sample range is* $7 - 3 = 4$.

Sampling Distribution

Definition

The probability distribution of a statistic is called a **sampling distribution**.

- Since a statistic is a random variable that depends only on the observed samples, it must have a probability distribution.
- The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples.

Sample Mean

- Suppose a random sample of n observations, $X_1, X_2, \ldots, X_n$, is taken from a *normal* distribution with mean μ and variance σ^2, i.e. $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$.
- The sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$$

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu$$

and variance

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}$$

Central Limit Theorem

Theorem

If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

- If $n \geq 30$, the normal approximation will be satisfactory regardless of population shape.
- If $n < 30$, the approximation is good only if the population is not too different from a normal distribution.
- If the population is known to be normal, the sampling distribution of $\bar{X}$ is normal for any size of n.

Central Limit Theorem

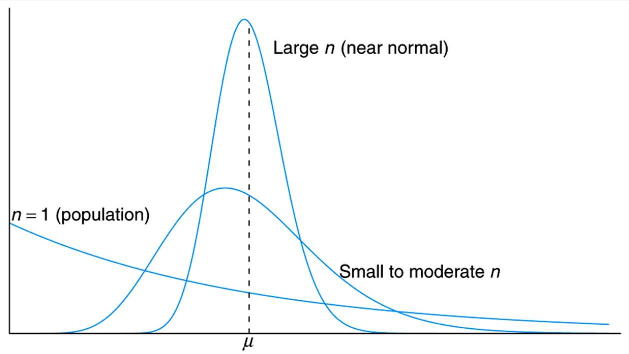


Figure 1: Illustration of CLT

Central Limit Theorem: Example

Example

An electronics company manufactures resistors that have a mean resistance of $\mu = 100\ \Omega$ and standard deviation $\sigma = 10\ \Omega$. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than $95\ \Omega$.

- The sampling distribution of $\bar{X}$ is normal, with mean $\mu_{\bar{X}} = 100\ \Omega$, and standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

- The probability $P(\bar{X} < 95) =$?

$$\begin{aligned} P(\bar{X} < 95) &= P\left(z = \frac{\bar{X} - 100}{2} < \frac{95 - 100}{2}\right) \\ &= P(z < -2.5) = 0.0062 \end{aligned}$$

Central Limit Theorem

Theorem (Two Samples of Two Populations)

If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{(\bar{X}_1 - \mu_1) - (\bar{X}_2 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

is approximately normal, when the conditions for the central limits theorem apply.

- The difference of two Gaussian RVs is still normal.
- If both n_1 and n_2 are greater than 30, the normal approximation of $X_1 - X_2$ is good.

Central Limit Theorem: Example

Example

The effective life of a part used in a jet engine is close to a normal random variable with mean 5000 hours and standard deviation 40 hours. An improvement has been introduced to increase the mean life to 5050 hours and to decrease the standard deviation to 30 hours. Suppose there are random samples of $n_1 = 16$ and $n_2 = 25$ components selected form the original and improved processes, respectively. What is the probability that the difference in the two sample means $\bar{X}_2 - \bar{X}_1$ is at least 25 hours.

- Assume that both sample means are normal.
- The RV $\bar{X}_1$ has mean 5000 hours and and standard deviation $\sigma_1/\sqrt{n_1} = 40/4 = 10$ hours.
- The RV $\bar{X}_2$ has mean 5050 hours and and standard deviation $\sigma_2/\sqrt{n_2} = 30/5 = 6$ hours.

Central Limit Theorem: Example

- The sample difference $\bar{X} = \bar{X}_2 - \bar{X}_1$ is also a normal random with mean $\mu_{\bar{X}} = \mu_{\bar{X}_2} - \mu_{\bar{X}_1} = 50$, and standard deviation

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_{\bar{X}_2}}{n_2} + \frac{\sigma^2_{\bar{X}_1}}{n_1} = 6^2 + 10^2 = 136$$

- The probability $P(\bar{x} = \bar{X}_2 - \bar{X}_1 \geq 25) =$?

$$\begin{aligned} P(\bar{x} \geq 25) &= P\left(z = \frac{\bar{X} - 50}{\sqrt{136}} \geq \frac{25 - 50}{\sqrt{136}}\right) \\ &= P(z \geq -2.14) \\ &= 1 - P(z < -2.14) \\ &= 0.9836 \end{aligned}$$

Two Samples: Example

Example (Paint Drying Time)

Two independent experiments are run in which two different types of paint are computed. Eighteen specimens are painted using type A, and drying time (in hours) is recorded each. The same is done with type B. The population standard deviations are both known to be 1.0. Assume that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_b$ are average drying times for samples of size $n_A = n_B = 18$.

- From the sampling distribution of $\bar{X}_A - \bar{X}_B$, we know that the sampling distribution is *approximately* normal.
- The mean is

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 0$$

Two Samples: Example

- The variance is

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{n} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}$$

- The probability of $P(\bar{X}_A - \bar{X}_B > 1.0)$ is

$$\begin{aligned} & P(\bar{X}_A - \bar{X}_B > 1.0) \\ = \quad & P\left((\bar{X}_A - \bar{X}_B) - \mu_{\bar{X}_A - \bar{X}_B} > 1.0 - 0.0\right) \\ = \quad & P\left(\frac{(\bar{X}_A - \bar{X}_B) - \mu_{\bar{X}_A - \bar{X}_B}}{\sigma_{\bar{X}_A - \bar{X}_B}} > \frac{1.0}{\sqrt{1/9}}\right) \\ = \quad & P(z > 3.0) = 1 - P(z < 3.0) \\ = \quad & 1 - 0.9987 = 0.0013 \end{aligned}$$

Two Samples: Example

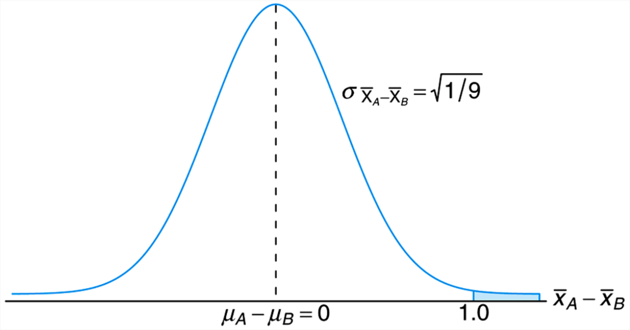


Figure 2: Area for $P(\bar{X}_A - \bar{X}_B > 1.0)$.

Sampling Distribution of Sample Variance I

- If a random sample of size n is drawn from a normal distribution with mean μ and variance σ^2. The sample variance (the statistic S^2) is given by

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

- It is easy to verify that

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}\left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&+ 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2
\end{aligned}$$

Sampling Distribution of Sample Variance II

- Dividing each term of the equality by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i - \bar{X})^2$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \quad = \quad \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ is a chi-squared random variable with n degrees of freedom.
- Since $\bar{X}$ is a normal distribution with mean μ and variance σ^2/n, the random variable $Z^2 = \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$ is also a chi-squared distribution with 1 degree of freedom.

Sampling Distribution of Sample Variance

Theorem

If S^2 is the variance of a random sample of size n taken from a normal populations having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

- The values of the random variable χ^2 are calculated from each sample by the formula:

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

- There is 1 less degree of freedom which is lost in estimation of μ, i.e., $\bar{X}$.

Chi-Squared Distribution

The probability that a random sample produces a χ^2 value greater than some specified value, α, is equal to the area under the curve to the right of this value.

$$P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$$

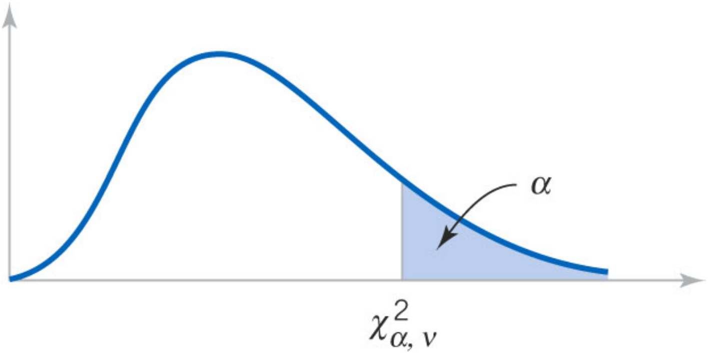
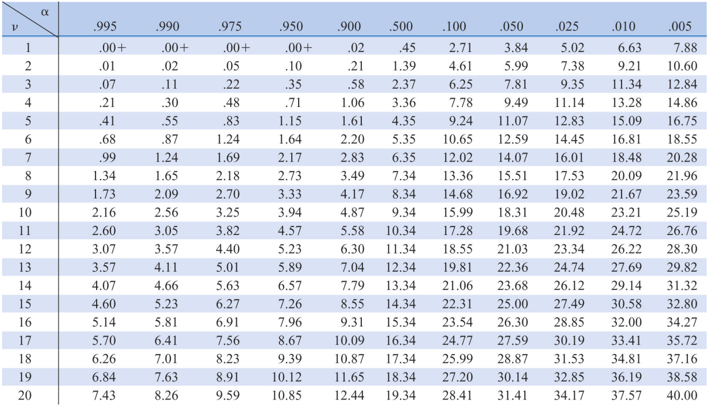


Figure 3: The chi-squared distribution with the degrees of freedom ν and the area of α.

Chi-Squared Distribution: Table

The following table gives values of χ^2_α for various values of α and ν.



α ν	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00

Figure 4: The chi-squared distribution with the degrees of freedom ν and the area of α.

Confidence Interval: Concept

- Exactly 95% of a chi-squared distribution lies between $\chi^2_{0.0975}$ and $\chi^2_{0.025}$.
- A χ^2 values falling to the right of $\chi^2_{0.025}$ is not likely to occur, $P < 0.025$, unless the assumed value of σ^2 is too small.
- A χ^2 values falling to the left of $\chi^2_{0.0975}$ is not likely to occur, $P < 0.025$, unless the assumed value of σ^2 is too small.
- When σ^2 is correct, it is possible, $P < 0.05$, to have a χ^2 value to the left of $\chi^2_{0.0975}$ or to the right of $\chi^2_{0.025}$.
- If this should happen, it is **more probable** that the assume value of σ^2 is in erro.

Sampling Distribution of Sample Variance: Example

Example

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation 1 year. If five of these batteries have lifetimes of $1.9, 2.4, 3.0, 3.5$ and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

- The sample variance is

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815$$

- The corresponding value of χ^2 is

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

with 4 degrees of freedom.

Sampling Distribution of Sample Variance: Example

- Since the 95% of the χ^2 values with 4 degrees of freedom fall between 0.484 and 11.143.
- Since $0.484 < 3.26 < 11.143$, the computed value is reasonable.
- The manufacturer has **no reason** to suspect the standard deviation is other than 1 year.

Chi-Squared Distribution

Theorem

If the random variable X is $N(\mu, \sigma^2)$, then the random variable $Y = \frac{(X-\mu)^2}{\sigma^2} = Z^2$ is $\chi^2(1)$.

- $Z = \frac{(X-\mu)}{\sigma}$ is $N(0,1)$.
- Since $f_Y(y) = \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})]$,

$$\begin{aligned} f_Y(y) &= \frac{1}{2^{1/2}\sqrt{\pi}} y^{1/2-1} e^{-y/2} \\ &= \frac{1}{\Gamma(\frac{1}{2})} \left(\frac{1}{2}\right)^{1/2} y^{1/2-1} e^{-y/2} \\ &= f_Y(y; \gamma = 1/2, \lambda = 1/2) = \frac{\lambda^\gamma x^{\gamma-1} e^{-\lambda x}}{\Gamma(\gamma)} \\ &= \chi^2(1) = \chi^2(\nu = 1) \end{aligned}$$

MGFs of Gamma Distributions

- If X is a Gamma distribution with pdf

$$f(x;\gamma,\lambda) = \begin{cases} \frac{\lambda^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-\lambda x}, & 0 < x < \infty \\ 0, & \text{elsewhere} \end{cases},$$

  then the MGF of the random variable X is

$$M_X(t) = \frac{1}{(1 - t/\lambda)^\gamma}$$

- The MGF of the chi-squared distribution with 1 degree of freedom , $\chi^2(1)$,is, i.e., $\lambda = 1/2$, $\gamma = \nu/2$ and $\nu = 1$

$$M_X(t) = \frac{1}{(1 - 2t)^{1/2}}$$

MGFs of Gamma Distributions

- If $X_1, X_2, \dots, X_n$ are independent random variables with moment generating functions $M_{X_1}(t), M_{X_2}(t), \dots, M_{X_n}(t)$, respectively, and if $Y = X_1 + X_2 + \cdots + X_n$ then the moment generating function of Y is

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \cdots \cdot M_{X_n}(t)$$

- If each $X_i, i = 1, \dots, n$ has a $\chi^2(1)$ distribution, then the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{n} \frac{1}{(1-2t)^{1/2}} = \frac{1}{(1-2t)^{n/2}}$$

which is also a chi-squared distribution with n degrees of freedom, $\chi^2(n)$.