Review of Last Lecture

Key concepts and/or techniques:

1. Sample mean: Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed with mean μ. Then the sample mean is defined as
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$
and a statistic and also an estimator of mean μ.

2. Mgf technique: Mgf, if exists, uniquely determines the distribution of the RV. Therefore, the distribution of a RV can be equivalently found via its mgf.

   Use the mgf technique to derive the distribution of
$$Y = \sum_{i=1}^{n} a_i X_i$$

Review of Last Lecture

[Theorem 5.4-1]

If $X_1, X_2, \cdots, X_n$ are independent RVs with respective mgfs $M_{X_i}(t)$ where $|t| < h_i$ for positive number h_i, $i = 1, 2, \cdots, n$. Then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t),$$

where $|a_i t| < h_i$, $i = 1, \cdots, n$.

Review of Last Lecture

[Theorem 5.4-2]

Let $X_1, X_2, \cdots, X_n$ be independent chi-square RVs with $r_1, r_2, \cdots, r_n$ degrees of freedom, respectively, i.e., $X_i \sim \chi^2(r_i), i = 1, \cdots, n$ Then

$$Y = X_1 + X_2 + \cdots + X_n \quad \text{is} \quad \chi^2(r_1 + r_2 + \cdots + r_n)$$

[Corollary 5.4-3]

If $X_1, X_2, \cdots, X_n$ are independent and have normal distributions $N(\mu_i, \sigma_i^2), i = 1, 2, ..., n$, respectively, then the distribution of

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n)$$

STA2001 Probability and Statistics I

Lecture 22

Tianshi Chen

The Chinese University of Hong Kong, Shenzhen

Section 5.5 Random function associated with normal distribution

Theorem 5.5-1, page 200

[Theorem 5.5-1]

If $X_1, X_2, \cdots, X_n$ are n independent normal variables with means $\mu_1, \mu_2, \cdots, \mu_n$ and variances σ_1^2, σ_2^2, $\cdots$, σ_n^2, respectively, then $Y = \sum_{i=1}^n a_i X_i$ has the normal distribution

$$Y \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Proof of Theorem 5.5-1, page 200

By Theorem 5.4-1, we have

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t) = \prod_{i=1}^{n} \exp(\mu_i a_i t + \frac{1}{2}\sigma_i^2 a_i^2 t^2)$$

$$= \exp\left\{ \left(\sum_{i=1}^{n} \mu_i a_i \right) t + \frac{1}{2} \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) t^2 \right\}$$

Example 1, page 201

Let X_1 and X_2 be the pounds of butter fat produced by 2 cows, respectively. Assume that

$$X_1 \sim N(693.2, 22820), X_2 \sim N(631.7, 19205)$$

and moreover, X_1 and X_2 are independent. What's the probability $P(X_1 > X_2)$?

Example 1, page 201

Let $Y = X_1 - X_2$. Then

$$Y \sim N(693.2 - 631.7, 22820 + 19205) = N(61.5, 42025)$$

$$P(X_1 > X_2) = P(Y > 0) = P\left(\frac{Y - 61.5}{\sqrt{42025}} > \frac{0 - 61.5}{\sqrt{42025}}\right)$$

$$= 1 - \Phi(-0.3) = 0.6179.$$

Corollary 5.5-1, Page 201

[Corollary 5.5-1]

If $X_1, X_2, \cdots, X_n$ is a random sample of size n from the normal distribution $N(\mu, \sigma^2)$, then the sample mean $\overline{X}$ has the following distribution

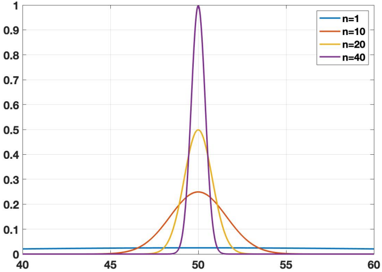
$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \Leftrightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Proof: Let $a_i = \frac{1}{n}, \mu_i = \mu, \sigma_i^2 = \sigma^2, i = 1, \cdots, n$. Then by Theorem 5.5-1, we obtain the result.

Example 2, page 201

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from $N(50, 16)$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(50, \frac{16}{n})$$



- To illustrate the effect of n: The larger n, the smaller the variance $\frac{16}{n}$.
- pdf of $\overline{X}$: The sharper the peak, the more concentrated in a small interval centered at 50.

Sample Variance

Definition

Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed with mean μ and σ^2. Then the sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

and an <u>estimator</u> of the variance σ^2, because

$$E(S^2) = \sigma^2.$$

Sample Variance

Note that

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(X_i-\overline{X}+\overline{X}-\mu\right)^2$$

$$=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)+\sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2$$

where

$$\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)=(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\overline{X})$$

$$=(\overline{X}-\mu)(\sum_{i=1}^{n}X_i-n\overline{X})=(\overline{X}-\mu)(n\overline{X}-n\overline{X})=0$$

Therefore,

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

Sample Variance

Then note that

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

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

Therefore,

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

Theorem 5.5-2, page 202

[Theorem 5.5-2]

Let $X_1, X_2, \cdots, X_n$ be random sample of size n from the normal distribution $N(\mu, \sigma^2)$ with $\sigma^2 > 0$. Then the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independent, and

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

The independence of $\overline{X}$ and S^2 is not proved here but deferred to Section 6.7 on page 294, and we only prove the second part.

Proof of Theorem 5.5-2, page 202

Following the proof of $E(S^2) = \sigma^2$, we have

$$\frac{n-1}{\sigma^2} S^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - \sum_{i=1}^{n} \left(\frac{\overline{X} - \mu}{\sigma} \right)^2$$

Now let

$$W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2, \quad Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Then

$$W = \frac{(n-1)S^2}{\sigma^2} + Z^2$$

Further note that $W \sim \chi^2(n), Z^2 \sim \chi^2(1)$, and moreover, S^2 and $\overline{X}$ are independent by assumption.

Proof of Theorem 5.5-2, page 202

Note that $W \sim \chi^2(n), Z^2 \sim \chi^2(1)$, S^2 and $\overline{X}$ are independent

$$E[e^{tw}] = E[e^{t(\frac{(n-1)S^2}{\sigma^2} + Z^2)}] = E[e^{t\frac{(n-1)S^2}{\sigma^2}}]E[e^{tZ^2}]$$

$$(1-2t)^{-\frac{n}{2}} = E[e^{t\frac{(n-1)S^2}{\sigma^2}}] \cdot (1-2t)^{-\frac{1}{2}}, \quad t < \frac{1}{2}$$

$$\Rightarrow E[e^{t\frac{(n-1)S^2}{\sigma^2}}] = (1-2t)^{-\frac{n-1}{2}}, \quad t < \tfrac{1}{2} \Rightarrow \tfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

A remark

Combining Corollary 5.4-3 and Thm 5.5-2 leads to the observation:

If $X_1, X_2, \cdots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n), \quad \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

When the mean μ is replaced by the sample mean $\overline{X}$, one degree of freedom is lost.

This is because there is an additional constraint

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

Example 5.5-3, page 204

Let X_1, X_2, X_3, X_4 be a random sample of size 4 from the normal

distribution $N(76.4, 383)$.

$$\sum_{i=1}^{4} \frac{(X_i - 76.4)^2}{383} \sim \chi^2(4), \quad \sum_{i=1}^{4} \frac{(X_i - \overline{X})^2}{383} \sim \chi^2(3)$$

Student's t Distribution

[Theorem 5.5-3]

Let

$$T = \frac{Z}{\sqrt{U/r}},$$

where $Z \sim N(0,1), U \sim \chi^2(r)$, and Z and U are independent. Then T has a student's t distribution

$$f(t) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi r}\Gamma(\frac{r}{2})} \frac{1}{(1+\frac{t^2}{r})^{\frac{r+1}{2}}}, \quad t \in (-\infty, \infty),$$

where r is called the degrees of freedom, and we simply write $T \sim t(r)$.

Sketch of the Proof of Theorem 5.5-3, page 204

Since Z and U are independent, their joint pdf $g(z,u)$ is

$$g(z,u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot \frac{1}{\Gamma(\frac{r}{2}) 2^{\frac{r}{2}}} u^{\frac{r}{2}-1} e^{-\frac{u}{2}}, \quad z \in R, u \in [0,\infty)$$

1. The cdf of T, $F(t)$ is

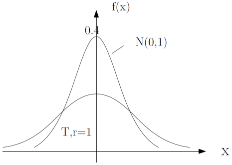
$$F(t) = P(T \le t) = P\left(\frac{Z}{\sqrt{\frac{U}{r}}} \le t \right) = P\left(Z \le \sqrt{\frac{U}{r}} t \right)$$

2. The pdf of T,

$$f(t) = F'(t)$$

Student's t Distribution: Heavy-tailed Distribution

Student's t distribution is a heavy tailed distribution



Standard normal distribution :

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{e^{\frac{1}{2}x^2}}, \quad x \in (-\infty, \infty)$$

Students' t distribution with $r = 1$:

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in (-\infty, \infty)$$

Therefore, Student's t distribution is a better choice than the normal distribution when the data contains outliers.

A Student's t RV based on Random Samples from Normal Distribution

By using the result of Corollary 5.5-1 and Theorems 5.5-2 and 5.5-3, we can construct an important student's t random variable.

Assume that $X_1, X_2, \cdots, X_n$ is a random sample of size n from a normal distribution $N(\mu, \sigma^2)$.

A Student's t RV based on Random Samples from Normal Distribution

Let

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \quad U = \frac{(n-1)S^2}{\sigma^2}$$

Then $Z \sim N(0,1)$ and $U \sim \chi^2(n-1)$. Since Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/(n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

A remark

If $X_1, \cdots, X_n$ is a random sample of size n from a normal distribution $N(\mu, \sigma^2)$, then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$