The χ^2 distribution & the t distribution

Lecture 12a (STAT 24400 F24)

CLT and normal distribution

Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean.

According to the central limit theorem, for large n,

$$\text{In distribution:} \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1), \qquad \bar{X} \approx \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

We can have probability statement, such as

$$\mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \le z_*\right) \approx 1-\alpha$$

For example,

$$\mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \le 1.96\right) \approx 95\%$$

Normal and t, χ^2 distributions

Note that, "$\mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \le 1.96\right) \approx 95\%$" is equivalent to

$$\mathbb{P}\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \approx 95\%$$

If σ^2 is known, this is a statement of a 95% confidence interval for μ.
However, in practice, σ^2 is not known.

How can we obtain a confidence interval by the CLT when σ^2 is unknown?
What if we replace σ^2 by sample variance S^2? (Lessons from Guinness)

In order to apply the CLT to learn about unknown parameters such as μ and σ^2, a few distributions (t, χ^2, F) closely related to the normal distribution are needed.

We start with the t distribution and the χ^2 distributions.

Sample mean & sample variance as estimators

Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2.

How could we estimate μ & σ^2 from the sample (i.e. the observed X_i's)?

The most common estimators are

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

Unbiasedness of the sample mean

These estimators are **unbiased**, meaning

$$\mathbb{E}(\bar{X}) = \mu, \qquad \mathbb{E}(S^2) = \sigma^2$$

For the sample mean, the unbiasedness can be shown by linearity of $\mathbb{E}(\cdot)$:

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

Unbiasedness of the sample variance

To show unbiasedness of the sample variance, we can calculate

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

this step is similar to $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

this explains dividing by n-1 instead of n

$$\Rightarrow \quad \mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^n \underbrace{\mathrm{Var}(X_i)}_{=\sigma^2} - \frac{n}{n-1}\underbrace{\mathrm{Var}(\bar{X})}_{=\sigma^2/n} = \sigma^2.$$

Normal sample mean & sample variance

Special case: normal distribution (i.e., $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$)

For normal data, it holds that

$$\bar{X} \perp\!\!\!\perp S^2$$

(even though they both are functions of the same sample)

<u>Proof for the case $n = 2$</u>: we have $\bar{X} = \frac{X_1+X_2}{2}$, and

$$S^2 = \frac{1}{2-1}\sum_{i=1}^2 (X_i - \bar{X})^2 = \left(X_1 - \frac{X_1+X_2}{2}\right)^2 + \left(X_2 - \frac{X_1+X_2}{2}\right)^2 = \frac{1}{2}(X_1 - X_2)^2$$

So, to show $\bar{X} \perp\!\!\!\perp S^2$, it's sufficient to check that $X_1 + X_2 \perp\!\!\!\perp X_1 - X_2$.

Normal sample mean & sample variance (cont.)

$$\begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{\text{linear transformation}} \underbrace{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}_{\text{bivariate normal}}$$

$\Rightarrow (X_1 + X_2, X_1 - X_2)$ is bivariate normal.

For components of a bivariate normal, uncorrelated $\iff$ independent.

$$\begin{aligned} &\mathrm{Cov}(X_1 + X_2, X_1 - X_2) \\ &\quad = \mathrm{Cov}(X_1, X_1) - \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_2, X_1) - \mathrm{Cov}(X_2, X_2) \\ &\quad = \mathrm{Var}(X_1) - 0 + 0 - \mathrm{Var}(X_2) = \sigma^2 - 0 + 0 - \sigma^2 = 0. \end{aligned}$$

Therefore, $X_1 + X_2 \perp\!\!\!\perp X_1 - X_2$ and so $\bar{X} \perp\!\!\!\perp S^2$.

function of $X_1 + X_2$ function of $X_1 - X_2$

The χ^2 distribution

If $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, then the distribution of $V = Z_1^2 + \cdots + Z_n^2$ is χ_n^2

(χ_n^2 — "the χ^2 distribution with n degrees of freedom")

Density:

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad \text{over} \quad x \geq 0$$

Expected value:

$$\mathbb{E}(V) = \mathbb{E}(Z_1^2) + \cdots + \mathbb{E}(Z_n^2) = \mathrm{Var}(Z_1) + \cdots + \mathrm{Var}(Z_n) = n.$$

Variance: $\mathrm{Var}(V) = 2n$. (exercise)

Note χ^2 is a special case of Gamma, with $\chi_n^2 = \mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$.

The χ^2 distribution - density

We can derive the density for the case $n = 1$.

Start with the CDF: for $x \geq 0$,

$$F_V(x) = \mathbb{P}(V \leq x) = \mathbb{P}(Z_1^2 \leq x) = \mathbb{P}(-\sqrt{x} \leq Z_1 \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x})$$

where Φ is the CDF of $N(0,1)$:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Take the derivative for density:

$$f_V(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_V(x) = \Phi'(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} - \Phi'(-\sqrt{x}) \cdot \frac{-1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}$$

$\Phi'(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ = the density of $\mathrm{N}(0,1)$

This is the density of $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$, $n = 1$ case of $\chi_n^2 \sim \mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$.

The χ^2 distribution

For normal data:

If $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2.$$

A more useful result:

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2 \qquad \Longrightarrow \qquad \frac{n-1}{\sigma^2} \cdot S^2 \sim \chi_{n-1}^2$$

Proof for the case $n = 2$: We know $\mathbb{E}(X_1 - X_1) = 0, \mathrm{Var}(X_1 - X_2) = 2\sigma^2$.

$$\frac{n-1}{\sigma^2} \cdot S^2 = \frac{1}{\sigma^2} \cdot \frac{1}{2} (\underbrace{X_1 - X_2}_{\sim \mathrm{N}(0, 2\sigma^2)})^2 = \left(\underbrace{\frac{X_1 - X_2}{\sqrt{2\sigma^2}}}_{\sim \mathrm{N}(0,1)} \right)^2 \sim \chi_1^2$$

Recap

What we've calculated so far:

If $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then

- $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (exact)
- $\frac{n-1}{\sigma^2} \cdot S^2 \sim \chi_{n-1}^2$ (exact)
- and $\bar{X} \perp\!\!\!\perp S^2$

If $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2 (not of normal distribution), then the above statements hold approximately, for large n.

The t distribution

If $Z \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_n$ and $Z \perp\!\!\!\perp V$,

then the distribution of $T = \frac{Z}{\sqrt{V/n}}$ is t_n,

"the t distribution with n degrees of freedom"

Density:

$$f(x) = (\text{normalizing constant}) \cdot \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad \text{over } x \in \mathbb{R}$$

which can be written as

$$f(x) \;\propto\; \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

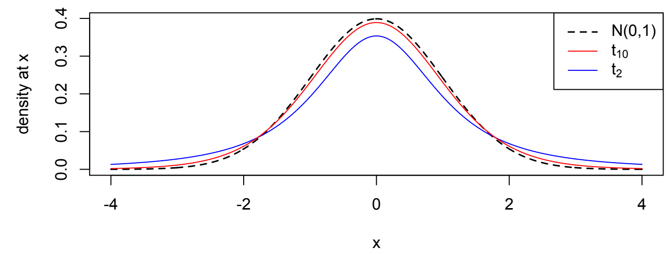
"$\propto$" means proportional to.

The degrees of freedom

For a small n, the t_n distribution has *heavy tails*:

$\mathbb{P}(T \geq x)$ is much larger than $1 - \Phi(x)$, as x grows large.

For increasing n, the t_n distrib. grows more similar to $\mathrm{N}(0,1)$.



We can see this in the density function:

$$f(x) \propto \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} = \left[\underbrace{\left(1 + \frac{x^2}{n}\right)^n}_{\approx e^{x^2}}\right]^{\overbrace{-(n+1)/2n}^{\approx -1/2}} \approx e^{-x^2/2}$$

t distribution for sampling

Return to the normal case: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Define

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Then:

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\overbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}^{\sim \mathrm{N}(0,1)}}{\sqrt{\underbrace{\frac{n-1}{\sigma^2} S^2}_{\sim \chi^2_{n-1}}/(n-1)}} \sim t_{n-1}$$

since numerator $\perp\!\!\!\perp$ denominator

Overview

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{n-1}{\sigma^2} \cdot S^2 \sim \chi^2_{n-1} \quad \text{and} \quad \bar{X} \perp\!\!\!\perp S^2$$

$$\text{and} \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2,

then the above statements hold approximately.

This important result is used to construct confidence intervals for μ.