MATH 4752/6752 – Mathematical Statistics II

Sampling Distributions

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

Let $f(\cdot;\theta)$ be the pdf of a specific distribution with unknown parameter θ.

Question: Can we estimate θ by getting samples of the iid RVs $X_1, \dots, X_n$ following $f(\cdot;\theta)$?

Definition. A set of iid RVs $X_1, \dots, X_n$ is called a **random sample** of their common distribution f. Given a specific function u, the random variable $Y = u(X_1, \dots, X_n)$ is called a **statistic**.

Example. Let $X_1, \ldots, X_n$ be iid RVs with pdf $f(\cdot; \theta)$. Then we can define two statistics:

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

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

Remark. In practice, we also apply the term of a statistic (e.g., sample mean and sample variance) to its actual value in an experiment.

Sampling distribution of the mean

Theorem. If $X_1, \ldots, X_n$ is a random sample of a distribution with mean μ and variance σ^2, then the sample mean $\bar{X}$ satisfies

$$\mathbb{E}[\bar{X}] = \mu, \qquad \text{var}[\bar{X}] = \frac{\sigma^2}{n}.$$

Proof. We can show that

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu,$$

$$\text{var}[\bar{X}] = \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}[X_i] = \frac{\sigma^2}{n}.$$

Remark. We often denote $\mathbb{E}[\bar{X}]$ by $\mu_{\bar{X}}$ and var$[\bar{X}]$ by $\sigma_{\bar{X}}^2$. Also $\sigma_{\bar{X}}$ is called the **sample error** of $\bar{X}$.

Theorem (Chebyshev's inequality). Let X be a RV with mean μ and variance σ^2, then for any $c > 0$, there is

$$\mathsf{P}(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Proof. We have that

$$\begin{aligned}
\mathsf{P}(|X-\mu| \geq c) &= \int_{|x-\mu|\geq c} f(x)\,dx \\
&\leq \int_{|x-\mu|\geq c} \frac{|x-\mu|^2}{c^2} f(x)\,dx \\
&\leq \int_{-\infty}^{\infty} \frac{|x-\mu|^2}{c^2} f(x)\,dx \\
&= \frac{\sigma^2}{c^2}.
\end{aligned}$$

Example. By Chebyshev's inequality, we have for any fixed $c > 0$ that

$$\mathsf{P}(|\bar{X} - \mu| \leq c) \geq 1 - \frac{\sigma^2}{n^2 c^2}.$$

Note that RHS tends to 1 as $n \to \infty$. This is informally known as the **Law of Large Numbers**.

Central Limit Theorem. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Denote $\bar{X}_n$ their sample mean. Define

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then the limiting distribution of Z_n as $n \to \infty$ is the standard normal distribution.

To prove CLT, we first recall several properties of MGFs.

Let $M_X(t)$ be the MGF of X and a, b be constants. Then

$$M_{X+a}(t) = \mathbb{E}[e^{(X+a)t}] = e^{at}\,\mathbb{E}[e^{Xt}] = e^{at} M_X(t),$$
$$M_{bX}(t) = \mathbb{E}[e^{bXt}] = \mathbb{E}[e^{X(bt)}] = M_X(bt),$$
$$M_{\frac{X+a}{b}}(t) = e^{\frac{at}{b}} M_{\frac{X}{b}}(t) = e^{\frac{at}{b}} M_X\Big(\frac{t}{b}\Big).$$

Proof of CLT. We notice that

$$M_{Z_n}(t) = M_{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}(t) = M_{\frac{n\bar{X}_n - n\mu}{\sqrt{n}\sigma}}(t) = e^{-\frac{\sqrt{n}\mu}{\sigma}t} M_{n\bar{X}_n}\Big(\frac{t}{\sqrt{n}\sigma}\Big).$$

Since $n\bar{X}_n = X_1 + \cdots + X_n$, we have

$$M_{n\bar{X}_n}\Big(\frac{t}{\sqrt{n}\sigma}\Big) = \prod_{i=1}^{n} M_{X_i}\Big(\frac{t}{\sqrt{n}\sigma}\Big) = \Big(M_X\Big(\frac{t}{\sqrt{n}\sigma}\Big)\Big)^n.$$

Also note that

$$M_X\Big(\frac{t}{\sqrt{n}\sigma}\Big) = 1 + \underbrace{\mu_1' \frac{t}{\sqrt{n}\sigma} + \frac{\mu_2'}{2}\Big(\frac{t}{\sqrt{n}\sigma}\Big)^2 + \cdots}_{=:\xi(t)}$$

where μ_i' is the ith moment of X. In particular, $\mu_1' = \mu$, $\mu_2' = \mu^2 + \sigma^2$.

Proof of CLT (cont). Recall that

$$\ln(1+x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots .$$

Hence we have

$$\begin{aligned}
\ln M_{Z_n}(t) &= -\frac{\sqrt{n}\mu}{\sigma}t + n \ln M_X\Big(\frac{t}{\sqrt{n}\sigma}\Big) \\
&= -\frac{\sqrt{n}\mu}{\sigma}t + n \ln(1+\xi(t)) \\
&= -\frac{\sqrt{n}\mu}{\sigma}t + n\Big(\xi(t) + \frac{\xi(t)^2}{2} + \cdots\Big) \\
&= \frac{t^2}{2} + \sum_{r=3}^{\infty} \frac{c_r t^r}{\sqrt{n^{r-2}}}
\end{aligned}$$

for constants c_r independent of t and n.

For any fixed $t \in (0, 1)$, we have

$$\sum_{r=3}^{\infty} \frac{c_r t^r}{\sqrt{n^{r-2}}} = O\Big(\frac{1}{\sqrt{n}}\Big) \to 0 \text{ as } n \to \infty.$$

Therefore ln $M_{Z_n}(t) \to \frac{t^2}{2}$, i.e., $M_{Z_n}(t) \to e^{t^2/2}$. This implies that the limiting distribution of Z_n is $N(0, 1)$, which proves CLT.

Remarks.

- It is $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$, not $\bar{X}_n$, that has density approaching that of the standard normal. When $n \geq 30$, the approximation accuracy is usually good enough.

- If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ for any $n \geq 1$.

Sample distribution with finite population and without replacement

Suppose we have a finite population $\{c_1, \dots, c_N\}$, and we select n of them in order without replacement. Let $X_1, \dots, X_n$ be the RVs representing our selections. Then the joint pmf of $(X_1, \dots, X_n)$ is

$$f(x_1, \dots, x_n) = \frac{1}{P_N^n} = \frac{(N-n)!}{N!}$$

The marginal distribution $f_r(x_r)$ of X_r is

$$f_r(x_r) = \sum_{x_s \neq x_r, s \neq r} f(x_1, \dots, x_n) = \frac{1}{P_N^n} \cdot P_{N-1}^{n-1} = \frac{(N-n)!}{N!} \cdot \frac{(N-1)!}{(N-n)!} = \frac{1}{N}$$

for any $x_r = c_1, \dots, c_N$.

To see the above, notice that when x_r is fixed, $(X_1, \dots, \widehat{x_r}, \dots, X_n)$ can take any permutation of the remaining $N-1$ objects (all but x_r).

For any $r = 1, \dots, n$, from the marginal pmf $f_r(c_r)$ we have

$$\mu_r = \mathbb{E}[X_r] = \sum_{i=1}^{N} c_i f_r(c_i) = \frac{1}{N} \sum_{i=1}^{N} c_i =: \mu$$

$$\sigma_r^2 = \mathbb{E}[(X_r - \mu)^2] = \sum_{i=1}^{N} (c_i - \mu_r)^2 f_r(c_i) = \frac{1}{N} \sum_{i=1}^{N} (c_i - \mu_r)^2 =: \sigma^2$$

For any $r \neq s$, the joint pmf of (X_r, X_s) is

$$g_{rs}(x_r, x_s) = \frac{1}{P_N^n} \cdot P_{N-2}^{n-2} = \frac{(N-n)!}{N!} \cdot \frac{(N-2)!}{(N-n)!} = \frac{1}{N(N-1)}$$

for any $x_r \neq x_s$.

From the joint pmf, we have

$$\begin{aligned}
\text{cov}(X_r, X_s) &= \mathbb{E}[(X_r - \mu)(X_s - \mu)] \\
&= \sum_{i \neq j} (c_i - \mu)(c_j - \mu) g_{rs}(c_i, c_j) \\
&= \sum_{i \neq j} (c_i - \mu)(c_j - \mu) \frac{1}{N(N-1)} \\
&= \frac{1}{N(N-1)} \sum_{i=1}^{N} (c_i - \mu) \sum_{j \neq i} (c_j - \mu) \\
&= -\frac{1}{N-1} \cdot \frac{1}{N} \sum_{i=1}^{N} (c_i - \mu)^2 \\
&= -\frac{1}{N-1} \sigma^2
\end{aligned}$$

where we used $\sum_{j \neq i}(c_j - \mu) = -(c_i - \mu)$ in the second last equality.

Now we can find the mean and variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$:

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu$$

$$\begin{aligned}\text{var}[\bar{X}_n] &= \text{var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \frac{1}{n^2}\text{var}[X_i] + 2\sum_{r<s}\frac{1}{n^2}\text{cov}(X_r, X_s) \\ &= n\cdot\frac{\sigma^2}{n^2} + \frac{n(n-1)}{2}\cdot\frac{2}{n^2}\cdot\left(-\frac{\sigma^2}{N-1}\right) = \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}\end{aligned}$$

Remark. We can see var$[\bar{X}_n]$ differs from $\frac{\sigma^2}{n}$ by a factor of $\frac{N-n}{N-1}$. If $N = n$, then there is no variance since $\bar{X}_n = \frac{1}{N}\sum_{i=1}^n c_i$ for sure. If $N \gg n$, then $\frac{N-n}{N-1} \approx 1$ which is close to the infinite population case.

Chi-square distribution

We have seen that if $Z \sim N(0,1)$, then $X := Z^2 \sim \Gamma(\frac{1}{2}, 2)$. Here X is said to have chi-square distribution with degree of freedom (df) 1. We denote $X \sim \chi_1^2$.

In general, X is said to have chi-square distribution with df ν if $X \sim \Gamma(\frac{\nu}{2}, 2)$, i.e.,

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2}$$

for $x > 0$ and $f(x) = 0$ if $x \le 0$. Hence

$$\mathbb{E}[X] = \frac{\nu}{2} \cdot 2 = \nu, \quad \text{var}[X] = \frac{\nu}{2} \cdot 2^2 = 2\nu, \quad M_X(t) = (1-2t)^{-\nu/2}.$$

Remark. Recall that if $X_i \sim \Gamma(\alpha_i, \beta)$ for $i = 1, \dots, n$ and are independent, then

$$Y = \sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{n} \alpha_i, \beta\right).$$

Therefore, if $Z_i \sim N(0, 1)$ are independent standard normal, then $Z_i^2 \sim \Gamma(\frac{1}{2}, 2)$ are independent χ_1^2, and

$$Y = \sum_{i=1}^{n} Z_i^2 \sim \Gamma\Big(\frac{n}{2}, 2\Big) = \chi_n^2.$$

Theorem. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be a random sample, then $\bar{X}$ and S^2 are independent, and

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

To prove this theorem, we need a series of lemmas.

Lemma. We have the following identities:

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

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

Lemma.

- If $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$.
- If $X_1, \ldots, X_n \sim N(0,1)$ is a random sample, then $Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$.

Lemma. If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ is a random sample, then $\bar{X}$ is independent of $X_i - \bar{X}$ for all $i = 1, \ldots, n$.

Sketch proof. The joint pdf of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2}.$$

Consider the transformation:

$$\begin{cases} Y_1 &= \bar{X} \\ Y_2 &= X_2 - \bar{X} \\ &\vdots \\ Y_n &= X_n - \bar{X} \end{cases} \quad \Longleftrightarrow \quad \begin{cases} X_1 &= Y_1 - Y_2 - \cdots - Y_n \\ X_2 &= Y_2 + Y_1 \\ &\vdots \\ X_n &= Y_n + Y_1 \end{cases}$$

Sketch proof (cont). Then the joint pdf of $Y_1, \dots, Y_n$ is

$$g(y_1, y_2, \dots, y_n) = C \cdot \underbrace{e^{-\frac{1}{2\sigma^2}((\sum_{i=1}^n y_i)^2 + \sum_{i=2}^n y_i^2)}}_{\text{fn of } y_2, \dots, y_n} \cdot \underbrace{e^{\frac{n}{2\sigma^2}(y_1 - \mu)^2}}_{\text{fn of } y_1}.$$

This implies that Y_1 is independent of $Y_2, \dots, Y_n$. Hence $\bar{X}$ is independent of $X_2 - \bar{X}, \dots, X_n - \bar{X}$ and thus also $X_1 - \bar{X} = -\sum_{i=2}^n (X_i - \bar{X})$.

With the lemma above, we can prove that $\bar{X}$ and S^2 are independent.

Proof of the theorem. Since $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$ is a function of $X_1 - \bar{X}, \ldots, X_n - \bar{X}$, we know $\bar{X}$ is independent of S^2.

Now recall that we have

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

Dividing σ^2 we obtain

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

Proof of the theorem (cont). Noticing that

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n, \qquad \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1,$$

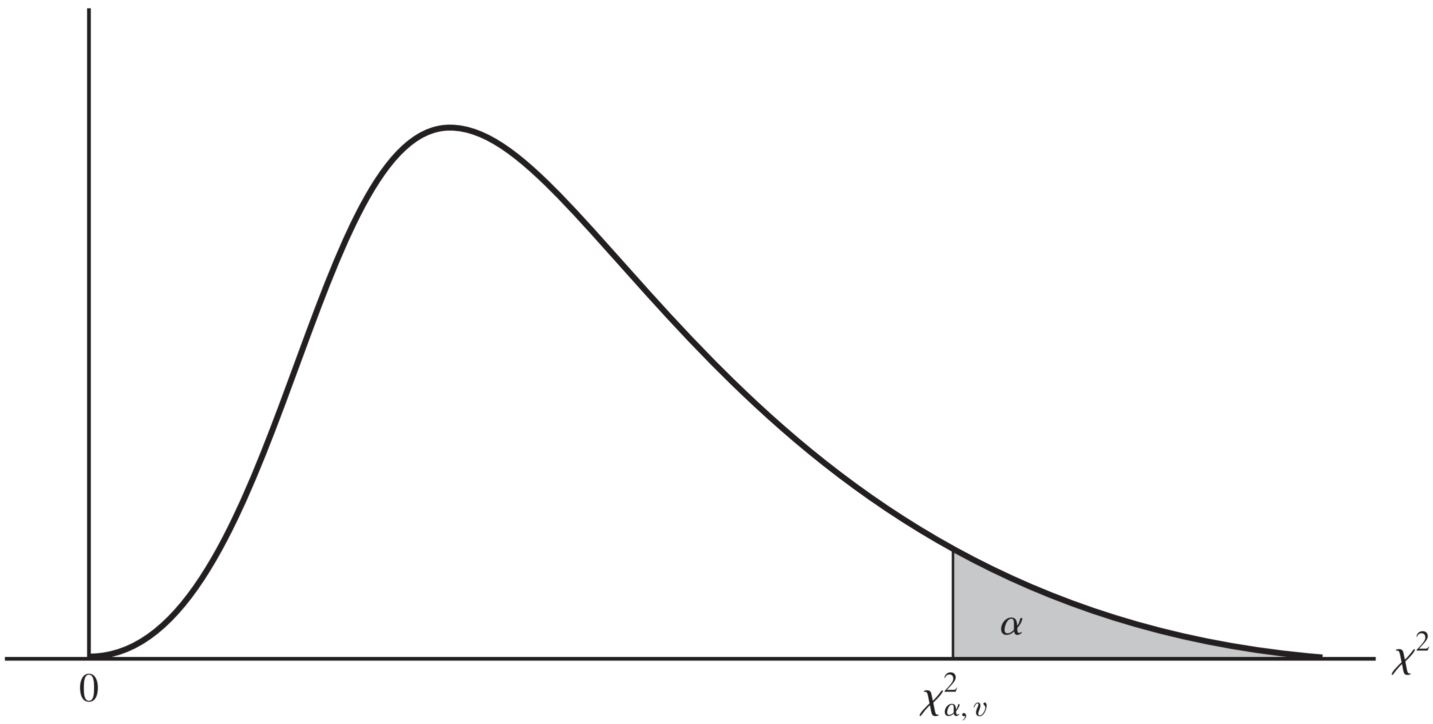
and that $\frac{(n-1)S^2}{\sigma^2}$ and $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ are independent, we get that

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

This completes the proof.

Let $X \sim \chi^2_\nu$ and $\alpha \in (0, 1)$, then $\chi^2_{\alpha,\nu}$ is the value such that

$$\mathsf{P}(X \geq \chi^2_{\alpha,\nu}) = \alpha$$



For certain given $\nu > 0$ and $\alpha \in (0, 1)$, we can look up the value of $\chi^2_{\alpha,\nu}$ in the χ^2 table (Table V in textbook):

Table V: Values of $\chi^2_{\alpha,\nu}$ †

ν	$\alpha = .995$	$\alpha = .99$	$\alpha = .975$	$\alpha = .95$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
1	.0000393	.000157	.000982	.00393	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548
7	.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. Suppose a semiconductor company wants to test the thickness of their semiconductors. They tested a sample of size 20 (assuming the thicknesses are from a normal distribution $N(\mu, \sigma^2)$). The production process is considered "out of control" if $\sigma > 0.60$ with probability 0.01. Suppose the test shows $s = 0.84$, is the process out of control?

Idea. Assuming $\sigma = 0.60$, we want to see how unlikely (i.e., with probability < 0.01) that $s = 0.84$ occurs. If it is indeed unlikely, we will declare that the assumption $\sigma = 0.60$ is inappropriate and we should have $\sigma > 0.60$.

Solution. The process is out of control if $\frac{(n-1)s^2}{\sigma^2}$ with $n = 20$ and $\sigma = 0.60$ exceeds $\chi^2_{0.01,19} = 36.191$. Since

$$\frac{(n-1)s^2}{\sigma^2} = \frac{19 \cdot (0.84)^2}{(0.60)^2} = 37.24 \ (> 36.191),$$

we declare that $\sigma = 0.60$ is inappropriate and the process is out of control.

The student t distribution

Suppose we have a random sample from a normal population $N(\mu, \sigma^2)$. Can we test the mean μ without knowing σ^2?

Theorem. Let $Y \sim \chi^2_\nu$ and $Z \sim N(0, 1)$ be independent, then

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has the probability density function given by

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < t < \infty$$

Here T is said to have the **student t distribution with df ν**, i.e., $T \sim t_\nu$.

Proof. First notice that the joint pdf of (Y, Z) is

$$f_{Y,Z}(y,z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \cdot \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\frac{\nu}{2}}} y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}.$$

Consider the transformation $(x,t) = \boldsymbol{u}(y,z)$ and its inverse $(y,z) = \boldsymbol{w}(x,t)$ where

$$(x,t) = \boldsymbol{u}(y,z) = \left(y, \frac{z}{\sqrt{y/\nu}}\right), \qquad (y,z) = \boldsymbol{w}(x,t) = \left(x, t\sqrt{x/\nu}\right).$$

So $\det(D\boldsymbol{w}(x,t)) = \sqrt{x/\nu}$.

Hence the joint pdf of (X, T) is

$$g(x,t) = f_{Y,Z}(\boldsymbol{w}(x,t))|\det(D\boldsymbol{w}(x,t))| = f_{Y,Z}(x, t\sqrt{x/\nu})\sqrt{x/\nu}$$

Applying the formula of $f_{Y,Z}$ and noticing that $Y = X$, we have

$$g(y,t) = \begin{cases} \frac{1}{\sqrt{2\pi\nu}\Gamma\left(\frac{\nu}{2}\right)2^{\frac{\nu}{2}}} y^{\frac{\nu-1}{2}} e^{-\frac{y}{2}\left(1+\frac{t^2}{\nu}\right)} & \text{for } y > 0 \text{ and } -\infty < t < \infty \\ 0 & \text{elsewhere} \end{cases}$$

For any fixed t, we notice that $g(y,t)$ is proportional to the pdf of $\Gamma(\alpha, \beta)$ where

$$\alpha = \frac{\nu+1}{2}, \quad \beta = \frac{2}{1+\frac{t^2}{\nu}}.$$

Hence we get

$$f_T(t) = \int_0^\infty g(y,t)\,dy = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < t < \infty.$$

Theorem. Suppose $\bar{X}$ and S^2 are respectively the sample mean and sample variance of a random sample from $N(\mu, \sigma^2)$, then

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

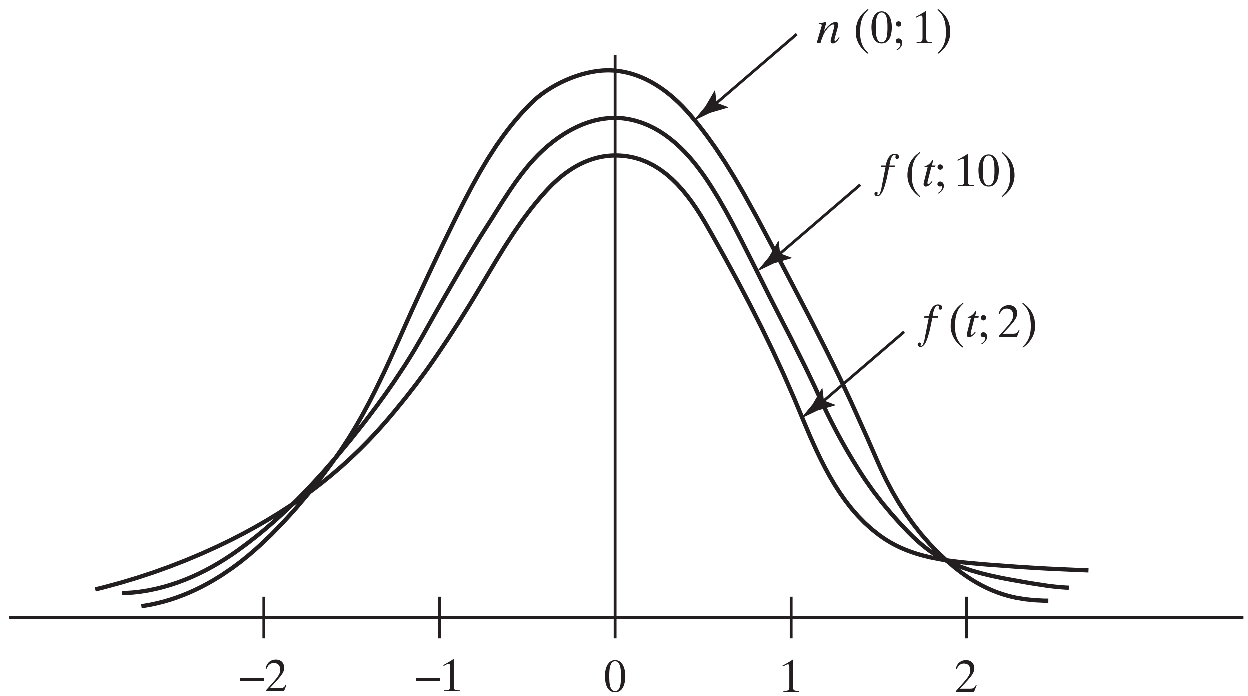
Proof. We let

$$Y := \frac{(n-1)S^2}{\sigma^2}, \qquad Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Then we know $Y \sim \chi^2_{n-1}$, $Z \sim N(0,1)$, and Y and Z are independent. Therefore,

$$T := \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} \sim t_{n-1}.$$

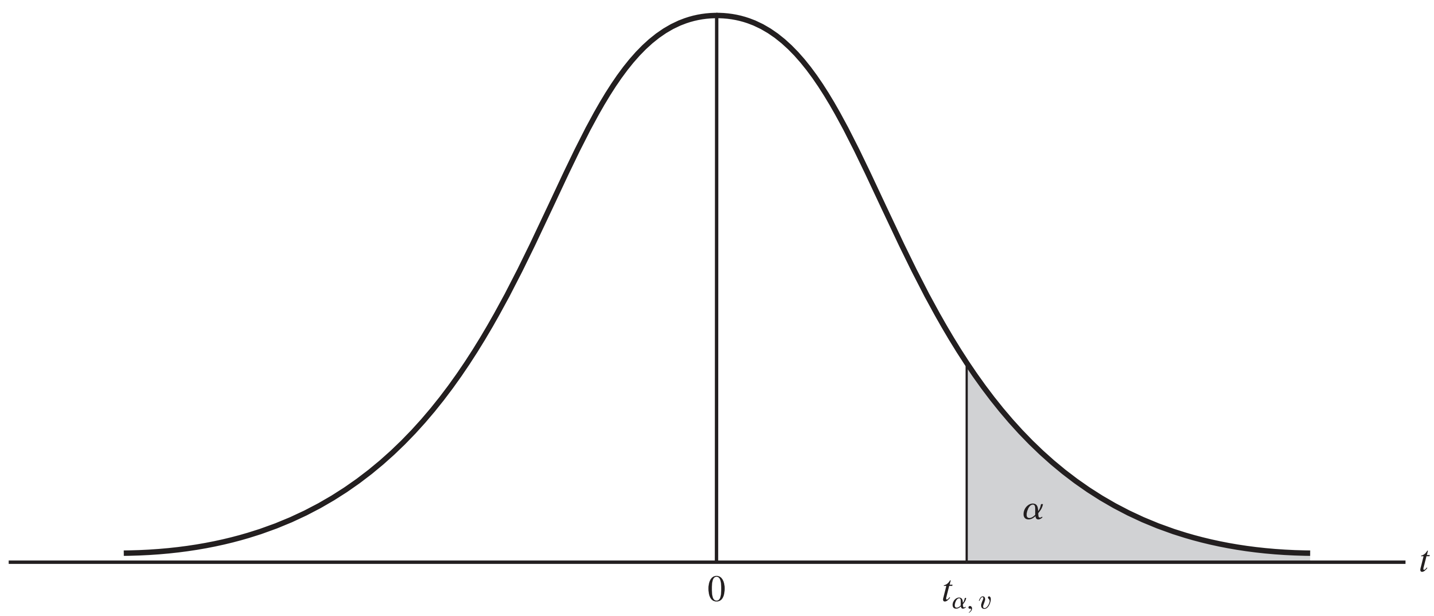
Comparison of the density functions of $N(0,1)$, t_2, and t_{10}:



Remark. t_ν is approximately $N(0,1)$ when $\nu \geq 30$.

Let $T \sim t_\nu$ and $\alpha \in (0.5, 1)$ (we do not need $\alpha \leq 0.5$ since f_T is symmetric about $t = 0$), then $t_{\alpha,\nu}$ is the value such that

$$\mathrm{P}(T \geq t_{\alpha,\nu}) = \alpha$$



For certain given $\nu > 0$ and $\alpha \in (0.5, 1)$, we can look up the value of $t_{\alpha,\nu}$ in the t-distribution table (Table IV in textbook):

Table IV: Values of $t_{\alpha,\nu}$ †

ν	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

Example. Suppose we obtain a random sample of size 16 from a normal population. Using this sample, we figure that $\bar{x} = 16.1$ and $s = 2.1$. Can we declare that the true mean $\mu > 12.0$ with confidence 0.99?

Idea. Assuming $\mu = 12.0$, we want to see how unlikely (i.e., with probability < 0.01) that $\bar{x} = 16.1$ occurs. If it is indeed unlikely, we will declare that the assumption $\mu = 12.0$ is inappropriate, and we should have $\mu > 12.0$.

Solution. Given that $n = 16$, $\bar{x} = 16.1$, $s = 2.1$, and assuming $\mu = 12.0$, we have

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{16.4 - 12.0}{2.1/\sqrt{16}} = 8.38.$$

On the other hand, we have $t_{0.005,15} = 2.947$ from the t-distribution table. Since $t \geq t_{0.005,15}$, we declare that the true mean $\mu > 12.0$ with confidence 0.99.

Fisher F distribution

Question: how do we draw statistical inferences about the ratio of two sample variances?

Theorem. Suppose $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ are independent, then

$$F = \frac{U/\nu_1}{V/\nu_2}$$

has the pdf given by

$$g(f) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot f^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} f\right)^{-\frac{1}{2}(\nu_1+\nu_2)}, & \text{if } f > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Here F is said to have the F**-distribution with degrees of freedoms** ν_1 **and** ν_2, denoted by $F \sim F_{\nu_1,\nu_2}$.

Proof. The joint pdf of (U,V) is

$$f_{U,V}(u,v) = \frac{1}{2^{\nu_1/2}\Gamma\left(\frac{\nu_1}{2}\right)} \cdot u^{\frac{\nu_1}{2}-1}e^{-\frac{u}{2}} \cdot \frac{1}{2^{\nu_2/2}\Gamma\left(\frac{\nu_2}{2}\right)} \cdot v^{\frac{\nu_2}{2}-1}e^{-\frac{v}{2}}$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \cdot u^{\frac{\nu_1}{2}-1}v^{\frac{\nu_2}{2}-1}e^{-\frac{\mu+v}{2}}$$

Consider the transformation $f = \frac{u/\nu_1}{v/\nu_2}$, then $u = \frac{\nu_1}{\nu_2}fv$ and hence $\frac{\partial u}{\partial f} = \frac{\nu_1}{\nu_2}v$. Thus the joint pdf of (F,V) is

$$g_{F,V}(f,v) = f_{U,V}\Big(\frac{\nu_1}{\nu_2}fv, v\Big) \cdot \frac{\nu_1}{\nu_2}v$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \cdot f^{\frac{\nu_1}{2}-1}v^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{v}{2}\left(\frac{\nu_1 f}{\nu_2}+1\right)}$$

for $f, v > 0$.

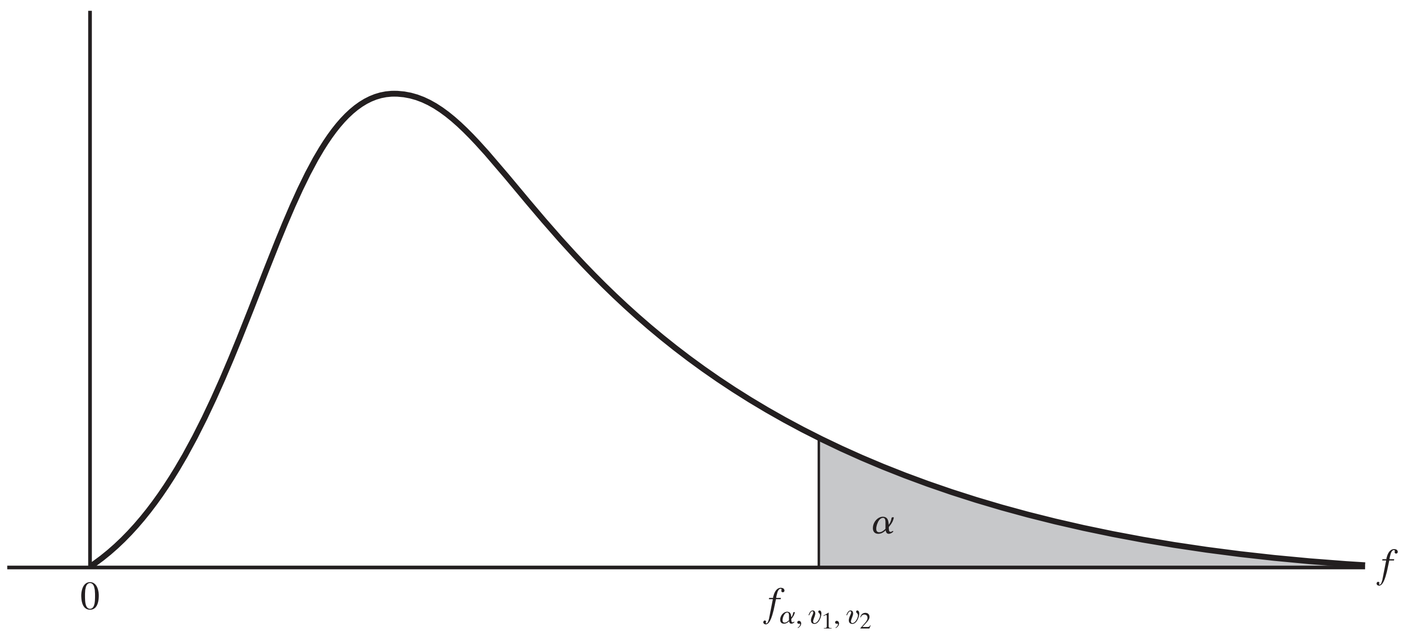
Integrating out v, we obtain the marginal pdf of F as

$$g(f) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot f^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} f\right)^{-\frac{1}{2}(\nu_1+\nu_2)}$$

for $f > 0$. It is obvious that $g(f) = 0$ if $f \le 0$.

Let $F \sim F_{\nu_1,\nu_2}$ and $\alpha \in (0, 1)$, then f_{α,ν_1,ν_2} is the value such that

$$\mathsf{P}(F \geq f_{\alpha,\nu_1,\nu_2}) = \alpha$$



For certain given $\nu_1, \nu_2 > 0$ and $\alpha \in (0, 1)$, we can look up the value of F_{α,ν_1,ν_2} in the F-distribution table (Table VI in textbook for $\alpha = 0.05$ and 0.01):

Table VI: Values of $f_{0.05,\nu_1,\nu_2}$ †

	ν_1 = Degrees of freedom for numerator																		
ν_2 = Degrees of freedom for denominator	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

Application of F statistics: compare the ratio of σ_1^2 and σ_2^2 from two independent normal populations.

Theorem. Suppose there are two independent normal populations with variances σ_1^2 and σ_2^2, and S_1^2 and S_2^2 are the sample variances of two random samples of size n_1 and n_2 from these two populations. Then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \sim F_{n_1-1,n_2-1}.$$

Proof. Notice that

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1}$$

for $i = 1, 2$ are independent.

Order Statistics

We consider nonparametric statistics (in contrast to parametric statistics before where we assumed normal population). Suppose $X_1, \dots, X_n \sim f$ is a random sample for an arbitrary f, then the **order statistics** are defined as

$$Y_1 = X_{(1)}, \quad Y_2 = X_{(2)}, \quad \dots, \quad Y_n = X_{(n)},$$

where $X_{(r)}$ is the r-th smallest one among $X_1, \dots, X_n$.

Question: what is the pdf of Y_r for $r = 1, \dots, n$?

Theorem. The pdf g_r of Y_r is given by

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)dx\right]^{n-r}$$

for $-\infty < y_r < \infty$.

Proof. For any $h > 0$, we partition $\mathbb{R}$ into three intervals using y_r and $y_r + h$, then the probability that $Y_1, \ldots, Y_{r-1}$ fall into the interval $(-\infty, y_r]$, Y_r falls into $(y_r, y_r + h]$, and $Y_{r+1}, \ldots, Y_n$ fall into $(y_r + h, \infty)$ is

$$\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)dx\right]^{r-1}\left[\int_{y_r}^{y_r+h} f(x)dx\right]\left[\int_{y_r+h}^{\infty} f(x)dx\right]^{n-r}.$$

If h is close to 0, then the probability above is $\mathrm{P}(y_r < Y_r \leq y_r + h)$ (since Y_{r+1} will be outside of this interval almost surely).

Proof (cont). On the one hand, we know

$$\frac{\mathsf{P}(y_r < Y_r \le y_r + h)}{h} = \frac{F_r(y_r + h) - F_r(y_r)}{h} \to g_r(y_r),$$

as $h \to 0$, where F_r is the cumulative distribution function of Y_r.

On the other hand, we have

$$\frac{1}{h}\int_{y_r}^{y_r+h} f(x)\,dx \to f(y_r)$$

$$\int_{y_r+h}^{\infty} f(x)\,dx \to \int_{y_r}^{\infty} f(x)\,dx$$

as $h \to 0$.

Combining the results above proves the theorem.

Several special order statistics

- **Minimal statistic** Y_1 has pdf

$$g_1(y_1) = n \cdot f(y_1) \left[\int_{y_1}^{\infty} f(x)dx \right]^{n-1} \quad \text{for } -\infty < y_1 < \infty$$

- **Maximal statistic** Y_n has pdf

$$g_n(y_n) = n \cdot f(y_n) \left[\int_{-\infty}^{y_n} f(x)dx \right]^{n-1} \quad \text{for } -\infty < y_n < \infty$$

- If $n = 2m + 1$ is odd, then the **sample median** Y_{m+1} has pdf

$$h(\tilde{x}) = \frac{(2m+1)!}{m!m!} \left[\int_{-\infty}^{\tilde{x}} f(x)dx \right]^{m} f(\tilde{x}) \left[\int_{\tilde{x}}^{\infty} f(x)dx \right]^{m}$$

for $-\infty < \tilde{x} < \infty$.

Example. Suppose $X_1, \ldots, X_n$ is a random sample from Exp(θ), i.e., the pdf is $f(x) = \frac{1}{\theta} e^{-x/\theta}$, then the pdf of Y_1 is

$$g_1(y_1) = \begin{cases} \frac{n}{\theta} \cdot e^{-ny_1/\theta} & \text{for } y_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The pdf of Y_n is

$$g_n(y_n) = \begin{cases} \frac{n}{\theta} \cdot e^{-y_n/\theta} \left[1 - e^{-y_n/\theta}\right]^{n-1} & \text{for } y_n > 0 \\ 0 & \text{elsewhere} \end{cases}$$

If $n = 2m+1$, then the pdf of the sample median Y_m is

$$h(\tilde{x}) = \begin{cases} \frac{(2m+1)!}{m!m!\theta} \cdot e^{-\tilde{x}(m+1)/\theta} \left[1 - e^{-\tilde{x}/\theta}\right]^{m} & \text{for } \tilde{x} > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Suppose f is continuous and nonzero at $\tilde{\mu}$ where $\tilde{\mu}$ is the **population median** such that

$$\int_{-\infty}^{\tilde{\mu}} f(x)\,dx = \frac{1}{2}.$$

Then for large $n = 2m + 1$, the sample median Y_m approximately follows the normal distribution:

$$N\Big(\tilde{\mu}, \frac{1}{4nf(\tilde{\mu})^2}\Big).$$

In particular, if $f(\cdot) = N(\cdot; \mu, \sigma^2)$ and sample size $n = 2m + 1$ is very large, then $f(\tilde{\mu}) = f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$ and there is approximately

$$Y_m \sim N\Big(\mu, \ \frac{\pi\sigma^2}{4m}\Big).$$

In contrast, the sample mean $\bar{X}_{2m+1} \sim N(\mu, \frac{\sigma^2}{2m+1})$ which has smaller variance.