

32. Sampling Distributions

Intro and Normal Distribution

χ^2 Distribution

t Distribution

F Distribution

Introduction and Normal Distribution

Goal: Talk about some distrn's we'll need later to do "confidence intervals" and "hypothesis tests".

Recall that a **statistic** is simply a function of the observations X_1, \dots, X_n from a random sample. The function does not depend explicitly on any unknown parameters.

Example: \bar{X} and S^2 are statistics.

Since statistics are RV's, it will sometimes be useful to figure out their distributions. The distribution of a statistic is called a **sampling distrn**.

Example: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{Nor}(\mu, \sigma^2/n)$.

The normal is often used to get confidence intervals for μ and to conduct hypothesis tests. Stay tuned!

We'll now introduce some other important sampling distrn's...

χ^2 Distribution

Definition/Theorem: If $Z_1, \dots, Z_k \stackrel{\text{iid}}{\sim} \text{Nor}(0, 1)$, then $Y \equiv \sum_{i=1}^k Z_i^2$ has the **chi-squared distrn with k degrees of freedom**.

Notation: $Y \sim \chi^2(k)$.

The p.d.f. is

$$f_Y(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} e^{-y/2}, \quad y > 0.$$

Fun Facts:

Can show that $E[Y] = k$ and $\text{Var}(Y) = 2k$.

The exponential distribution is a special case. In fact,
 $\chi^2(2) \sim \text{Exp}(1/2)$.

For $k > 2$, the $\chi^2(k)$ p.d.f. is skewed to the right.
(You get an occasional “large” observation.)

Recall that $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the gamma fn.

Tables for **quantiles** of the χ^2 can be found in the back of the book.

Definition: The $(1 - \alpha)$ quantile of a RV X is that value x_α such that $P(X > x_\alpha) = 1 - F(x_\alpha) = \alpha$. Note that $x_\alpha = F^{-1}(1 - \alpha)$, where $F^{-1}(\cdot)$ is the **inverse c.d.f.** of X .

Plenty of upcoming examples will make this concept clear.

Notation: If $Y \sim \chi^2(k)$, then we denote the $(1 - \alpha)$ quantile with the special symbol $\chi_{\alpha,k}^2$ (instead of x_α). In other words, $\Pr(Y > \chi_{\alpha,k}^2) = \alpha$.

Example: If $Y \sim \chi^2(10)$, then

$$\Pr(Y > \chi_{0.05,10}^2) = 0.05,$$

where we can look up $\chi_{0.05,10}^2 = 18.31$.

Another Property: χ^2 's add up. If Y_1, \dots, Y_n are *indep* with $Y_i \sim \chi^2(d_i)$, $\forall i$, then $\sum_{i=1}^n Y_i \sim \chi^2(\sum_{i=1}^n d_i)$.

Proof: Just use m.g.f.'s. Won't go thru it here.

So where does the χ^2 distrn come up in statistics?

It usually arises when we try to estimate σ^2 .

Example: If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, then we'll show in the next module that

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1} \sim \frac{\sigma^2 \chi^2(n-1)}{n-1}.$$

t Distribution

Definition/Theorem: Suppose that $Z \sim \text{Nor}(0, 1)$, $Y \sim \chi^2(k)$, and Z and Y are indep. Then $T \equiv Z/\sqrt{Y/k}$ has the **Student's t distrn with k degrees of freedom.**

Notation: $T \sim t(k)$. Further, the p.d.f. is

$$f_T(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k} \Gamma\left(\frac{k}{2}\right)} \left(\frac{x^2}{k} + 1\right)^{-\frac{k+1}{2}}, \quad x > 0.$$

Fun Facts:

The $t(k)$ looks like the $\text{Nor}(0,1)$, except the t has fatter tails.

In fact, as the degrees of freedom k becomes large, $t(k) \rightarrow \text{Nor}(0, 1)$.

Can show that $E[T] = 0$ and $\text{Var}(T) = \frac{k}{k-2}$ ($k > 2$).

Notation: If $T \sim t(k)$, then we denote the $(1 - \alpha)$ quantile by $t_{\alpha,k}$. In other words, $\Pr(T > t_{\alpha,k}) = \alpha$.

Example: If $T \sim t(10)$, then

$$\Pr(T > t_{0.05,10}) = 0.05,$$

where we find $t_{0.05,10} = 1.813$ in the back of the book.

So what do we use the t distribution for in statistics?

It's used when we find confidence intervals and conduct hypothesis tests for the mean μ . Again, stay tuned.

By the way, why did I originally call it **Student's** t distrn?

“Student” is the pseudonym of the guy who originally derived it.

F Distribution

Definition/Theorem: Suppose that $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, and X and Y are indep. Then $F \equiv mX/nY$ has the F **distrn with n and m degrees of freedom.**

Notation: $F \sim F(n, m)$. Further, the p.d.f. is

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{x^{\frac{n}{2}-1}}{\left(\frac{n}{m}x + 1\right)^{\frac{n+m}{2}}}, \quad x > 0.$$

Fun Facts:

The $F(n, m)$ is usually a bit skewed to the right.

Note that you have to specify two d.f.'s.

Can show that $E[F] = \frac{m}{m-2}$ ($m > 2$) and $\text{Var}(F) =$
blech.

t distribution is a special case — can you figure out which?

Notation: If $F \sim F(n, m)$, then we denote the $(1 - \alpha)$ quantile by $F_{\alpha, n, m}$. I.e., $\Pr(F > F_{\alpha, n, m}) = \alpha$.

Tables are in the back of the book for various α, n, m .

Example: If $F \sim F(5, 10)$, then

$$\Pr(F > F_{0.05, 5, 10}) = 0.05,$$

where we find $F_{0.05, 5, 10} = 3.33$ in the back of the book.

Remark: It can be shown that $F_{1-\alpha, m, n} = 1/F_{\alpha, n, m}$.

Use this fact if you have to find something like $F_{0.95, 10, 5} = 1/F_{0.05, 5, 10} = 1/3.33$.

So what do we use the F distribution for in statistics?

It's used when we find confidence intervals and conduct hypothesis tests for the ratio of variances from two different processes. Details later.