1 Sampling Distributions

In the first semester we focused on Probability. This semester we will focus on the foundations of Statistics, which will make use of the probability that we learned last semester.

The core goal of statistics is to describe a particular characteristic of a population.

- Usually the population is very large and getting a clear picture of the entire population would be difficult and expensive (i.e. the average height of uconn students, the median household income in the state of CT)
- We instead perform an action called inference
  - Collect a sample, a small portion of the whole population
  - Evaluate the characteristic of interest of the sample
  - Extrapolate to the larger population
- In order to perform inference we will need to sample from the population. This sample is often referred to as the data.
  - Each item in a sample that we measure in some way can be thought of as a random variable
  - We assume that all of the items in a sample are independent from each other
  - From a sample we calculate a statistic that we then use to perform our inference

Definition 1. *A **Statistic** is a function of observable random variables and known constants.*

We can think of a statistic as something that takes data and turns it into a number. Because the statistic is a function of random variables, it is itself, a random variable with a distribution. We often call this distribution a sampling distribution

1.1 Sampling Distributions

Common Statistics and their Sampling distributions:
One common situation that calls for statistics is when a sample is collected from a population that is assumed to be Normally distributed with a mean of μ and a variance of σ^2. For example suppose you asked 100 UConn alumni how much student debt they had upon graduating from school (assuming they had more than none); it is not unreasonable to assume the distribution of the debt held by UConn alumni is approximately normally distributed, and so each alumnus that you spoke to would be represented as an observation from that normal distribution. Furthermore, assuming a reasonable sampling procedure, it would be typical to expect that the amount of debt one alumnus had would be independent of the amount of debt that another had. This means you can

think of your sample of 100 alumni as independent and identically distributed (i.i.d.) $N(,\mu,\sigma^2)$.

In a situation like this where we have a sample of n i.i.d. observations (represented using the random variables $X_1,\ldots,X_n$) from a Normal distribution of mean μ and variance of σ^2, then the sample mean and sample variance as defined below have the following distributions

Theorem 1 (**Sample Mean of i.i.d.** $N(\mu,\sigma^2)$**).** *Let $X_1,X_2,X_3,...,X_n$ be i.i.d. $N(\mu,\sigma^2)$, then $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean, and is distributed $N(\mu,\frac{\sigma^2}{n})$.*

Proof. Let $U=\sum_{i=1}^n \frac{X_i}{n}(=\bar{X})$.

$$\begin{aligned}
M_U(t) &= E[e^{Ut}] \\
&= E[e^{(\sum_{i=1}^n \frac{X_i}{n})t}] \\
&= E[e^{(\sum_{i=1}^n \frac{X_i}{n}t)}] \\
&= E[e^{\frac{t}{n}X_1}e^{\frac{t}{n}X_2}...e^{\frac{t}{n}X_n}] \\
&= E[e^{\frac{t}{n}X_1}]E[e^{\frac{t}{n}X_2}]...E[e^{\frac{t}{n}X_n}] \leftarrow \text{since we assume that the items in the sample are independent} \\
&= M_{X_1}(\frac{t}{n})M_{X_2}(\frac{t}{n})...M_{X_n}(\frac{t}{n}) \\
&= e^{\mu\frac{t}{n}+\frac{(t/n)^2}{2}\sigma^2}e^{\mu\frac{t}{n}+\frac{(t/n)^2}{2}\sigma^2}...e^{\mu\frac{t}{n}+\frac{(t/n)^2}{2}\sigma^2} \leftarrow \text{ n terms} \\
&= e^{n\mu\frac{t}{n}+\frac{(t/n)^2}{2}n\sigma^2} \\
&= e^{\mu t+\frac{t^2}{2}\frac{\sigma^2}{n}}
\end{aligned}$$

Thus $U=\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$

□

Theorem 2 (**Sample Variance of i.i.d.** $N(\mu,\sigma^2)$**).** *Let $X_1,X_2,X_3,...,X_n$ be i.i.d. $N(\mu,\sigma^2)$, then $S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$ is the sample variance and it can be shown that $\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\bar{X})^2\sim\chi^2_{n-1}$*

Theorem 3 (**Independence of the sample mean and sample variance of i.i.d.** $N(\mu,\sigma^2)$**).** *Let $X_1,X_2,X_3,...,X_n$ be i.i.d. $N(\mu,\sigma^2)$, then $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ (the sample mean) and $S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$ (the sample variance) are independent*

Remark. *Because the sample variance has a scaled χ^2 distribution, it will be helpful to know that since, $\bar{X}$ and S^2 are independent, transformations of $\bar{X}$ and S^2 are also independent. Luckily, this independence is generally true, and is stated in general in the following theorem:*

Theorem 4 (**Independence of transformed independent Random Variables).** *] Let X and Y be two random variables that are independent and let $g(\cdot)$ and $h(\cdot)$ be one-to-one functions (i.e. their functional inverses exist). Then, $g(X)$ and $h(Y)$ will also be independent.*

Additionally, there are other sampling distributions that are the result of combinations of independent random variables:

Definition 2. *Let Z be a standard normal random variable (i.e. $Z \sim N(0,1)$), W be a random variable with a χ^2_n distribution, and let Z and W be independent. Then*

$$T = \frac{Z}{\sqrt{W/n}}$$

is said to have a t_n distribution with n degrees of freedom.

Definition 3. *Let W_1 and W_2 both have a χ^2 distribution with n_1 and n_2 degrees of freedom, respectively. Additionally, let W_1 and W_2 be independent. Then,*

$$F = \frac{W_1/n_1}{W_2/n_2}$$

is said to have an F_{n_1,n_2} distribution with n_1 and n_2 degrees of freedom

1.2 Examples

Suppose that you work for a drug company who is working with a new drug that (hopefully) improves test scores on Statistics exams. To analyze this, you've set up an experiment where you test 4 students' abilities by giving them a statistics exam(out of 100), give them the drug and then test them again with another exam that covers the same material, but is different (still out of 100). For now, we will assume that the distribution of the differences between the scores before and after taking the drug for each student is Normal with a mean of 0 and a variance of 1.

- We can think of each each student's difference in score as a random variable with a $N(0,1)$ distribution
- $\Rightarrow X_1, X_2, X_3, X_4$ each represent a student's difference in exam scores, and each is distributed $N(0,1)$
- $\Rightarrow \bar{X} = \frac{1}{4}\sum_{i=1}^{4} X_i \sim N(0, \frac{1}{4})$
- $\Rightarrow 2\bar{X} \sim N(0,1)$
- $\Rightarrow 3S^2 = \sum_{i=1}^{4}(X_i - \bar{X})^2 \sim \chi^2_3$
- $\Rightarrow \frac{2\bar{X}}{\sqrt{S^2}} \sim t_3$

1.3 Exercises

Suppose you also have a sample of 9 students $(Y_1, ..., Y_9)$ perform the same experiment the other 4 students did, but using a different drug produced by a competing company. Suppose the distribution of differences in the exam scores for this other drug) are known to be normally distributed with a mean of 1 and a variance of 4.

What is the distribution of

1. $\bar{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i$?
2. $2S_Y^2 = \frac{8}{4}\frac{1}{8}\sum_{i=1}^{9}(Y_i - \bar{Y})^2$?
3. $\frac{Y_1-1}{2}$?
4. $\frac{\sum_{i=1}^{9}\frac{Y_i-1}{2}}{9}$?
5. $\frac{1}{3}\sum_{i=1}^{9}\frac{Y_i-1}{2}$?
6. $\frac{\frac{1}{3}\sum_{i=1}^{9}\frac{Y_i-1}{2}}{\sqrt{2S^2/8}}$?
7. $\frac{S_X^2}{S_Y^2/4}$, where $S_X^2 = \frac{1}{3}\sum_{i=1}^{4}(X_i - \bar{X})^2$?

1.4 Solutions

1. $Y_1, ..., Y_9$ are all distributed $N(1, 4)$ and are assumed independent
   $\Rightarrow \bar{Y} \sim N(1, \frac{4}{9})$
2. For the same reason, we know that $2S_Y^2 \sim \chi_8^2$
3. $$\begin{aligned} Y_1 \sim N(1,4) &\Rightarrow Y_1 - 1 \sim N(0,4) \\ &\Rightarrow \frac{Y_1 - 1}{2} \sim N(0,1) \end{aligned}$$
4. From the problem above we see that $\frac{Y_i-1}{2} \sim N(0,1)$ for all i.
   $\Rightarrow \frac{\sum_{i=1}^{9}\frac{Y_i-1}{2}}{9} \sim N(0, \frac{1}{9})$
5. From the problem above we can deduce that:
$$\begin{aligned} \frac{1}{3}\sum_{i=1}^{9}\frac{Y_i-1}{2} &= \frac{3}{9}\sum_{i=1}^{9}\frac{Y_i-1}{2} \\ &\Rightarrow \frac{3}{9}\sum_{i=1}^{9}\frac{Y_i-1}{2} \sim N(0,1) \end{aligned}$$

6. From # 2 we know that $2S_Y^2 \sim \chi_8^2$ and from the previous problem we know that $\frac{3}{9}\sum_{i=1}^{9}\frac{Y_i-1}{2} \sim N(0,1)$. Therefore

$$\frac{\frac{1}{3}\sum_{i=1}^{9}\frac{Y_i-1}{2}}{\sqrt{2S^2/8}} \sim t_8$$

7. From # 2 we know that $2S_Y^2 \sim \chi_8^2$ and from the example we know that $3S_X^2 \sim \chi_3^2$. Therefore

$$\frac{S_X^2}{S_Y^2/4} \sim F_{3,8}$$