Chapter 8: Sample Statistics

Pranav Tikkawar

October 1, 2024

Definition: A random sample of size n from a population with pdf $f(x)$ is a sequence of n independent random variables with pdf $f(x)$.
Thus $X_1, X_2, \ldots, X_n$ are independent random variables with pdf $f(x)$.
Example: X_i = amount of ice cream in the ith scop with the same scoop
Question: What can we infer about the distribution Sample must be diret to the joint pdf
eg: $P(X_1 > X_2 + X_3)$
The jpdf of X_1, X_2, X_3 is $f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3)$

$$P(X_1 > X_2 + X_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{x_1=x_2+x_3}^{\infty} f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3$$

Integral over the region $\mathbb{R}^3$ **Definition** A statistic is a random var which is a funtion of the random sample
Example: Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Theorem: Suppose $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance σ^2. Then $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$
Theorum Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with distribution $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
Proof: Idea get MGF of $\bar{X}$

$$\begin{aligned} M_{\bar{X}}(t) &= M_{1/n\sum X_i}(t) \\ &= M_{\sum x_i}(t/n) \\ &= M_{X_1}(t/n)^n \end{aligned}$$

We know $M_N(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

$$M_{X_1}(t/n)^n = e^{\mu t + \frac{\sigma^2 t^2}{2n}}$$

Suppose X is a rv. Consider $P(|X - \mu_X| < k\sigma_X) \geq 1 - 1/k^2$ **Theorem:** Chebyshev's Inequality
Proof:

$$P(|X - \mu_X|^2 < k^2\sigma_X^2) = \int_{\mu-k\sigma}^{\mu+k\sigma} f(x)dx$$

Application:

$$P(|\bar{X}-\mu|<k\sigma)\geq 1-\frac{1}{k^2}$$
$$=P(|\bar{X}-\mu|<k\sigma/\sqrt{n})\geq 1-\frac{1}{k^2}$$
$$\rightarrow P(|\bar{X}-\mu|<\tilde{k})\geq 1-\frac{\sigma^2_{pop}}{n\tilde{k}^2}$$

If X is a rv with finite nonzero variance σ^2, then fixing an interval around μ,

$$\sigma^2=E[(X-\mu)^2]=-\int_{-\infty}^{\infty}|X-\mu^2|f(x)\ dx=\int_{near}\cdots+\int_{far}\cdots$$
$$=\int_{\mu-k}^{\mu+k}\cdots+\int_{X:|X-\mu|\geq k}\cdots$$

Since integrand is non-negative because $|x-\mu^2|\geq 0$ and $f(x)\geq 0$, then the first term drops out to create inequality,

$$\sigma^2\geq\int_{|X-\mu|\geq k}|x-\mu^2|f(x)\ dx$$

since $|x-\mu^2|\geq k^2$,

$$\sigma^2\geq k^2\int_{|X-\mu|\geq k}f(x)\ dx$$
$$\frac{\sigma^2}{k^2}\geq P(|X-\mu|\geq k)$$

Chebyshev's Inequalities

$$P(|X-\mu|\geq k)\leq\frac{\sigma^2}{k^2}\iff P(\text{outside})\text{ is bounded above}$$
$$P(|X-\mu|<k)\geq 1-\frac{\sigma^2}{k^2}\iff P(\text{inside})\text{ is bounded below}$$

Applying to $\overline{X}$ gives "Weak Law of Large Numbers"(W-LLN),

$$P\left(|\overline{X}-\mu<k|\right)\geq 1-\frac{\sigma^2}{n_k^2}$$

Since $\mu_{\overline{X}}=\mu$ and $\sigma^2_{\overline{X}}/k^2=\sigma^2/nk^2$.

Q:

How large should n be so that $\overline{X}$ approx's μ_{pop} with error less than 10^{-2} with prob. > 0.99? $\sigma_{pop} = 0.2$

A:

Using W-LLN,

$$\begin{aligned} P(|\overline{X} - \mu| < 10^{-2}) \geq 1 - \frac{0.2^2}{n(10^{-}2)^2} &\geq 0.99 \\ 0.01 &\geq \frac{0.04}{10^{-4}n} \\ n &\geq 40,000 \end{aligned}$$

Note: error is a statistic because its a rv that depends on random sample.

Central Limit Theorem:

Suppose $X_1, \ldots, X_n$ is a random sample iid from a pop. with well-def mgf. Then the dist of standardized $\overline{X}$ approaches *standard normal*.

$$P\left(a \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right)$$

since $\mu_{\overline{X}} = \mu$. As $n \to \infty$,

$$P(a \leq Z \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.$$

Rmk: to standardize a rv A means to subtract mean and divide by std. dev,

$$\begin{aligned} B = \frac{A - \mu_A}{\sigma_A} &\to E[B] = \frac{1}{\sigma}(E[A] - \mu_A E[1]) = 0 \\ &\to V[B] = \frac{1}{\sigma_A^2} V[A - \mu_A] = \frac{V[A]}{\sigma_A^2} = 1 \end{aligned}$$

Q:

It's known that amt of ice cream in 1 scoop is a rv which follows an unknown distribution with mean $\mu = 2$g, $\sigma = 0.1$g. Find an approx. for the prob that after $n = 100$ scoops, a total of more than 200.02g.

A:

Let X_i = amt in ith scoop. The event is $X_1 + \cdots + X_{100} \geq 200.02$. Using that $\overline{X} = \sum_{i=1}^{100} X_i/100$, standardizing, and CLT,

$$P\left(\frac{\overline{X}-2}{0.1/10} \geq \frac{2.0002-2}{0.1/10}\right) \approx P(Z \geq 0.02)$$

Application to Bernoulli

$$X \sim \text{Ber}(p) = \begin{cases} 1 & \text{with prob p} \\ 0 & \text{with prob 1-p} \end{cases}$$

Apply CLT to $X_1, \ldots, X_n$ iid Ber(p),

$$\frac{\frac{\sum_{i=1}^n X_i}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}}$$

above has distribution approaching Z as $n \to \infty$. Take $\sum_{i=1}^n X_i$ is sum of n indp Ber rv's as rv Y with Bin(n,p). So 1 binomial rv Y,

$$\frac{Y-np}{\sqrt{np(1-p)}} \sim Z$$

where Z is standard normal.
Sample Statistic
We looked at $\overline{X}$ so far.
We want to define and explore Sample Variance statistic.

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

Thus $\Gamma(1/2) = \int_0^\infty t^{-1/2}e^{-t}dt$

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty t^\alpha e^{-t}dt \\ &= \left[-t^\alpha e^{-t}\right]_0^\infty + \alpha\int_0^\infty t^{\alpha-1}e^{-t}dt \\ &= \alpha\Gamma(\alpha)\end{aligned}$$

We say X is a Gamma r.v w/ parameters $\alpha, \beta > 0$ if its pdf is

$$f(X) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, x > 0 \\ 0, \text{otherwise} \end{cases}$$

Question: PDF of $Y = Z^2$ where $Z \sim N(0,1)$
Note: $Y \geq 0$
Answer: $P(0 \leq Y \leq y) = P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y})$

$$
\begin{aligned}
&= P(Z^2 \leq y) \\
&= P(-\sqrt{y} \leq Z \leq \sqrt{y}) \\
&= 2P(0 \leq Z \leq \sqrt{y}) \\
&= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz
\end{aligned}
$$

This is the CDF (cumulative distribution function) of Y.

$$
\begin{aligned}
f_Y(y) &= \frac{d}{dy} F(x) \\
&= \frac{d}{dy} 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \frac{2}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} \\
&= \frac{1}{\sqrt{2\pi y}} e^{-y/2}
\end{aligned}
$$

Maria notes:

- $\overline{x}$: sample mean statistic
- want to define and explore sample variance statistic

Gamma fn:

for $\alpha > 0$, the following is a Gamma fn,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt$$

If $\alpha < 1 \rightarrow$ vertical asymptote

- near $t = 0$ is still ok because $\int_0^\infty \frac{1}{sqrtt} \, dt$ is defined (p-integral, take lower bound as r and evaluate as $r \rightarrow 0$)

When $\alpha = 1$,

$$\Gamma(1) = \int_0^\infty e^{-t} \, dt = \left(-e^{-}t\right|_0^\infty = 1$$

When $\alpha = \alpha + 1$,

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty t^\alpha e^{-t}\ dt,\\ &= \left(t^\alpha e^{-t}\right|_0^\infty - \int_0^\infty \alpha t^{\alpha-1}\cdot -e^{-t}\ dt,\\ &= 0 + \alpha\Gamma(\alpha),\end{aligned}$$

by IBP where $u = t^\alpha$, $du = \alpha t^{\alpha-1}$, $v = -e^{-t}$, and $dv = e^{-t}dt$. So, if $\boxed{\alpha > 0, \Gamma(\alpha+1) = \alpha\Gamma(\alpha)}$. Further, if $n \in \mathbb{Z}_{>0}$, then $\Gamma(n) = (n-1)!$.

- **Gamma dist:** X is a gamma RV with parameters $\alpha > 0$, $\beta > 0$ if its pdf is,

$$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)},\ x > 0$$

- $\beta^\alpha\Gamma(\alpha)$ is the normalization factor.

Calculation of normalization factor:

$$\begin{aligned}\int_0^\infty x^{\alpha-1}e^{\frac{-x}{\beta}}\ dx &= \int_{u=0}^{u=\infty} \beta^{\alpha-1}u^{\alpha-1}e^{-u}\beta\ du,\\ &= \beta^\alpha \cdot \Gamma(\alpha)\end{aligned}$$

by u-sub with $u = x/\beta$, $dx = \beta du$.

- gamma with $\alpha = 1$: has dist. $exp(\lambda = \beta) = \frac{1}{\beta}e^{-x/\lambda}$

Q

pdf of $Y = Z^2$? where $Z \sim N(\mu = 0, \sigma^2 = 1)$

A

Can first find cdf of Y. Since $y \geq 0$,

$$\begin{aligned}P(0 \leq Y \leq y) = P(Z^2 \leq y) = P(-\sqrt(y) \leq Z \leq \sqrt(y)) &= 2\cdot P(0 \leq Z \leq \sqrt{y}),\\ &= 2\int_0^{\sqrt{y}} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\ dz\end{aligned}$$

$P(-\sqrt(y) \leq Z \leq \sqrt(y)) = 2\cdot P(0 \leq Z \leq \sqrt{y})$ because Z has symmetry. Calculating pdf from cdf of Y,

$$\frac{d}{dy}\frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{y}} e^{-z^2/2}\ dz = \frac{1}{2\sqrt{y}}\frac{2e^{-\left(\sqrt{y}^2\right)/2}}{\sqrt{2\pi}}$$

So,

$$f_Y(y) = \frac{e^{-y/2}}{y^{1/2}\sqrt{2\pi}},\ y > 0$$

Note: this is the pdf of Gamma with $\alpha = 1/2$, $\beta = 2$ because $\Gamma(1/2) = \sqrt{\pi}$.

- **def (Chi-Square):** X has a Chi-Square (χ^2_ν) with $\nu > 0$ degrees of freedom if it is a Gamma rv with parameters $\alpha = \nu/2$ and $\beta = 2$.
- so, dist of Z^2 is $\chi^2_{\nu=1}$

Moments of Gamma

$$\begin{aligned}
\mu'_r = E[X^r] &= \int_0^\infty \frac{x^r x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\ dx, \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{r+\alpha-1} e^{-x/\beta}\ dx, \\
&= \frac{\beta^{r+\alpha}}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty u^{r+\alpha-1} e^{-u}\ du,
\end{aligned}$$

where $x = u\beta$, $dx = \beta du$. The integral above is the same as $\Gamma(1+\alpha)$,

$$E[X^r] = \frac{\beta^r}{\Gamma(\alpha)} \Gamma(r+\alpha)$$

Expectaion of X^r is:

$$\begin{aligned}
E[X^r] &= \int_0^\infty \frac{x^r x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\ dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{r+\alpha-1} e^{-x/\beta}\ dx \\
&= \frac{\beta^{r+\alpha}}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty u^{r+\alpha-1} e^{-u}\ du \\
&= \frac{\beta^r}{\Gamma(\alpha)} \Gamma(r+\alpha)
\end{aligned}$$

Thus μ is $\beta\alpha$ and second moment is $\beta^2\alpha(\alpha+1)$.
Thus the variance of X is $\beta^2\alpha$.
Exponential

$$E[exp(\lambda)] = \lambda$$

$$Var[exp(\lambda)] = \lambda^2$$

chi-square

$$E[\chi^2_\nu] = \nu$$

$$Var[\chi^2_\nu] = 2\nu$$

MGF will be

$$\sum_{n=0}^{\infty} \frac{\mu_r' t^r}{r!}$$

$$= \sum_{n=0}^{\infty} \frac{\alpha(\alpha+1)\dots(\alpha+r-1)\beta^r t^r}{r!}$$

This is $(1-\beta t)^{-\alpha}$

Sample Variance Statistic

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

Important identity:

$$\sum (X_i - \bar{X})^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Lets say want $E[S^2]$.
using the definition will not fully work because we dont know $E[(X_i - \bar{x})^2]$
We can use the identity above to get $E[S^2]$

$$E[S^2] = \frac{1}{n-1}(\sum E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2])$$

The first term is the expectation of the sample pop squared.
The second term is the variance of the sample mean.

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

Thus the expectation of the sample variance is the population variance.
Theorum: $X_1 ... X_n$ is a random sample from a normal pop with mean μ and variance σ^2. Then
a) $\bar{X}$ and S^2 are independent
b) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
Aside: Proof of :

$$\sum (X_i - \bar{X})^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$= n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Since S^2 is a statistic (a random variable) its good to have its pdf (in terms of pop pdf) we dont answer in general but we do for a normal population.
It has a gamma distribution with $\alpha = \frac{\nu}{2}$ and $\beta = 2$
This is also known as a chi-square distribution with ν degrees of freedom.

So its a chi-square distribution with ν degrees of freedom.
We can also see see that
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
Prove this using the fact that the population is normal.
Proof: Each of the X_i is normal with mean μ and variance σ^2
The left hand side become
$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$= \sum (X_i - \bar{X})^2/\sigma^2$$
$$= \sum (X_i - \mu)^2/\sigma^2 - n(\bar{X} - \mu)^2/\sigma^2$$
We can define $Z_i = \frac{X_i - \mu}{\sigma}$
Then Z_i is standard normal with $\mu = 0, \sigma^2 = 1$. because X_i is normal
Let $\tilde{Z} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
Then $\tilde{Z}$ is standard normal
Since we proved earlier that if each X_i is normal then $\bar{X}$ is normal
Thus $\tilde{Z}$ is standard normal
$$\frac{(n-1)S^2}{\sigma^2} + \tilde{Z}^2 = \sum_{i=1}^{n} Z_i^2$$
$$\tilde{Z}^2 \sim \chi^2_1$$
$$\sum Z_i^2 \sim \chi^2_n$$
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
$\frac{(n-1)S^2}{\sigma^2}$ and $\tilde{Z}^2$ are independent prove this
We learned that
$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

0.1 8.7 Order Statistics

Given random sample $X_1, X_2, \ldots, X_n$
the rth order statistic Y_r has rhe value that is the rthvalue when the sample is ordered from smallest to largest.
So $r = 1, 2, \ldots, n$
Example:
Suppose $X_1 = 3, X_2 = \pi, X_3 = e$
Then $Y_1 = e, Y_2 = \pi, Y_3 = 3$
Note Y_1 is also called the sample minimum and Y_n is called the sample maximum.

Sample Median is the middle one. If n is odd, it is the middle value. If n is even, it is the average of the two middle values.
We actually know the pdf of the order statistics.
Fix an interval $[a, b]$

$$P(Y_r \in [a,b]) = P(a \leq Y_r \leq b)$$

$$P(\text{once of the X's is in } [a,b] \text{and r-1 before and n-r after})$$

$$= \frac{n!}{(r-1)!(n-r)!}\int_a^b f(x)dx(\int_{-\infty}^a f(x)dx)^{r-1}(\int_b^\infty f(x)dx)^{n-r}$$

The first term is the combination of the elements of the sample: aka the multinomial coefficient: $\binom{n}{r-1,1,n-r}$
The first integral is the probability that one of the X's is in the interval
The second integral is the probability that r-1 of the X's are before the interval
The third integral is the probability that n-r of the X's are after the interval
This proability is $\int_a^b f_{Y_r}(y_r)dy_r$
So let $a = y_r$ and $b = y_r + h$

$$\lim_{h\to 0}\frac{n!}{(r-1)!(n-r)!}\int_{y_r}^{y_r+h}\frac{f(x)}{h}dx(\int_{-\infty}^{y_r} f(x)dx)^{r-1}(\int_{y_r+h}^\infty f(x)dx)^{n-r}$$

$$f_{Y_r}(y_r) = \frac{n!}{(r-1)!(n-r)!}f(y_r)(\int_{-\infty}^{y_r} f(x)dx)^{r-1}(\int_{y_r}^\infty f(x)dx)^{n-r}$$

Now in general for a uniform distribution, the pdf of the rth order statistic is

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

$$f_{Y_r}(y_r) = \frac{n!}{(r-1)!(n-r)!}\frac{1}{b-a}\frac{y_r-a}{b-a}^{r-1}\frac{b-y_r}{b-a}^{n-r}$$

This is applicable for Y_r in $[a, b]$

$$f_{Y_r}(y_r) = \frac{n!}{(r-1)!(n-r)!}\frac{(y_r-a)^{r-1}(b-y_r)^{n-r}}{(b-a)^n}$$

Example $n = 3, r = 1$,

$$f_{Y_1}(y_1) = 3\frac{(y_1-a)^0(b-y_1)^2}{(b-a)^3} = \frac{3(b-y_1)^2}{(b-a)^3}$$

Question: Y_1 in an exponetional population with pdf

$$f(x) = \frac{e^{-x/\lambda}}{\lambda}$$

What is the pdf of Y_1?
Answer: $n = n, r = 1$

$$f_{Y_1}(y_1) = \frac{n!}{(r-1)!(n-r)!}\frac{e^{-y_1/\lambda}}{\lambda}(\int_{y_1}^{\infty}\frac{e^{-x/\lambda}}{\lambda}dx)^{n-1}$$

$$= n\frac{e^{-y_1/\lambda}}{\lambda}(e^{-y_1/\lambda})^{n-1}$$

$$= n\frac{e^{-ny_1/\lambda}}{\lambda}$$

We can recognize this as an exponential distribution with parameter λ/n