Variance

$$Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

Tossing a coin

$$E[X] = 0 * (1-p) + 1 * p = p$$

$$E[X^2] = E[X] = p$$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

Sample variance

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

More samples → the better concentration around the population val

---

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \sigma^2/n \leftarrow \text{variance of sample mean}$$

$$\underset{\text{sample variance}}{S^2} \xrightarrow{\text{proxy}} \underset{\text{population variance}}{\sigma^2}$$

- Logical estimate is s^2/n
- Logical estimate of the standard error is $S/\sqrt{n}$

S, sd, talks about how variable the population is

$S/\sqrt{n}$, se, talks about how variable averages of random samples of size n from the population are.

Standard uniforms have variance $1/12$ means of random samples of n uniforms have sd $1/\sqrt{12 \cdot n}$

Poisson $\rightarrow$ sd: $2/\sqrt{n}$

Common distributions

PMF : $P(X = x) = p^x (1-p)^{1-x}$ $\boxed{\text{Bernoulli}}$

Let $X_1 \ldots X_n$ be Bernoulli (p)

then $X = \sum_i^n X_i$ is a binomial random value

Binomial MF

$$p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$\binom{n}{x}$ is called "n choose x" $= \dfrac{n!}{x!(n-x)!}$

$\binom{n}{0} = \binom{n}{n} = 1$

Ex. friend has 8 children, 7 girls

Each sample has an independent probability for each birth, what is the probability of getting 7 or more girls out of 8 births?

$$\binom{8}{7} 0{,}5^7 (1-0{,}5)^1 + \binom{8}{8} 0{,}5^8 (1-0{,}5)^0 \simeq 0{,}04$$

Normal distribution

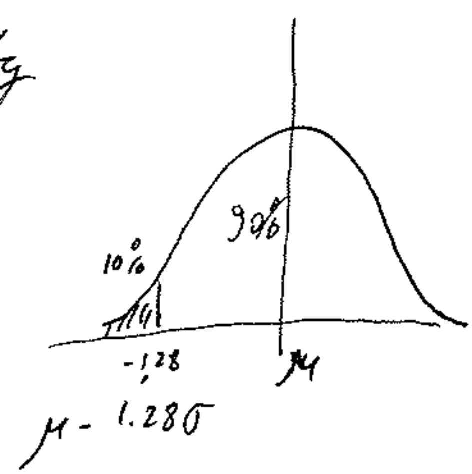
$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

$$X \sim N(\mu, \sigma^2)$$

facts about normal density

if $X \sim N(\mu, \sigma^2)$ then

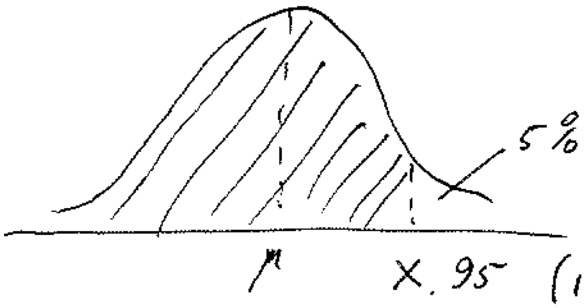
$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$



if Z is standard normal

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

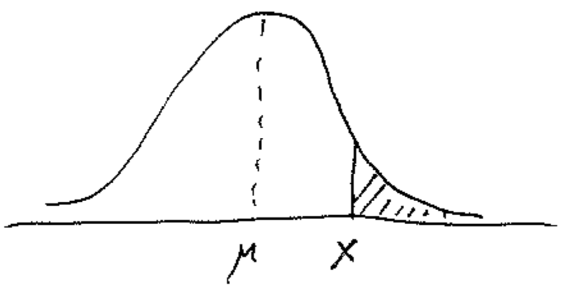
what is the 95th percentile of a $N(\mu, \sigma^2)$?



(1,645 standard)

R: qnorm (0.95, mean = mu, sd = sd)

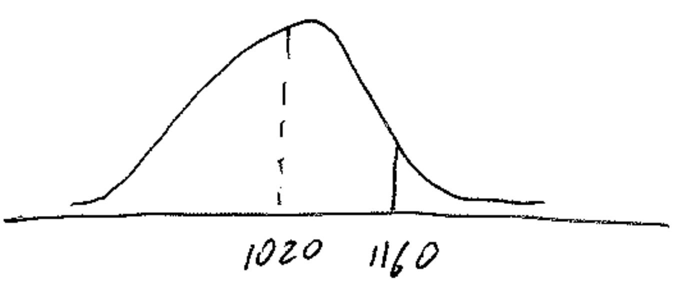
$X_{.95} \rightarrow \mu + \sigma\, 1{,}645$



$$\frac{X - \mu}{\sigma}$$

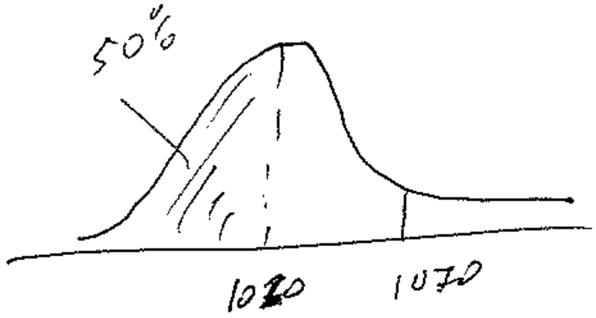
Ex. daily ad clicks $\sim N(\mu, \sigma^2)$, $\mu = 1020$, $sd = 50$

what is the probability of getting more than 1.160 clicks in a day?



2.8

Ex. what number of daily ads would represent the one where 75% days would have fewer clicks?



qnorm (0.75, mean = 1020, sd=50)

Poisson

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

mean is λ
Var is λ
↑ must be equal.

Modeling count data
Model event-time or survival data
Model contigency data
Approximating binomials when n large and p small

Poisson is used to model rates

$X \sim Poisson(\lambda t)$ where $\lambda = E[X/t]$ is expected count per unit of time
$t \to$ total monitoring time

Ex. people at bus stop is Poisson with a

mean of 2.5 per hour

Time of observation $\rightarrow$ 4 hours

What is the probability that 3 or

fewer people show up for the whole time

ppois (3, lambda = 2,5 · 4)

Poisson approximation for binomial

n is large

p is small

$X \sim$ Binomial (n, p)

$\lambda = np$

pbinom (2, size = 500, prob = 0.01)

ppois (2, lambda = 500 · 0.01)

Asymptotics

Limits of random variables

Law of large numbers says that the average limits at what its estimating, the population mean

Ex. $\bar{X}_n$ average of the result n coin flips, proportion of heads

An estimator is consistent if it converges to what you want to estimate.
If we collect infinite number of samples, we will get the exact number, which is the population mean.
SD and var are consistent as well.

Central limit Theorem

CLT states that the distribution of averages of iid values becomes that of a standard normal as the sample size increases.

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate - Mean of estimate}}{\text{Std. Err. of estimate}}$$

Ex.

Let X_i be the outcome of die i

Then note that $\mu = E[X_i] = 3.5$

$Var(X_i) = 2.92$

$SE \sqrt{2.92/n} = 1.71/\sqrt{n}$

Roll n dice, take their mean, substract off 3.5 and divide by $1.71/\sqrt{n}$

Let X_i be the 0 or 1 result of the i^{th} flip of a possibly unfair coin

- The sample proportion, say $\hat{p}$, is the average of the coin flips.

$E[X_i] = p$ and $Var(X_i) = p(1-p)$

$SE \rightarrow \sqrt{p(1-p)/n}$

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \rightarrow \sim N(\mu, \sigma^2)$$

Confidence intervals

$\bar{X}$ is approx N with μ and sd $\sigma/\sqrt{n}$

probably $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 5%

$\bar{X} \pm 2\sigma/\sqrt{n}$ is called a 95% interval for μ

Sample proportions

In the event that each X_i is 0 or 1 with common success probability p then $\sigma^2 = p(1-p)$

The interval takes form

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Replacing p with $\hat{p}$ in the se results is what is called a Wald confidence interval for p

For 95% intervals

$$\hat{p} \pm \frac{1}{\sqrt{n}}$$

Ex random sample of 100 likely voters, 56 intent to vote for you.

- Can you relax
- How persize is the estimate?

1/sqrt(100) = 0.1   CI = c(0.46, 0.66)

Not enough to relax, do more!

round(1/sqrt(10^(1:6)), 3)

Quick fix for the CI simulation (see code)

n is not large enough for the CLT to be applicable for many of the values of p

Quick fix: ~~wrappa~~ the interval with.

$$\frac{X + 2}{n + 4}$$

Add two successes failures, Agresti/Coull interval

Poisson interval

A nuclear pump failed 5 times out of 94.32 days, give a confidence interval for the failure rate per day.

$X \sim \text{Poisson}(\lambda t)$

Estimate $\hat{\lambda} = X/t$

$\text{Var}(\hat{\lambda}) = \lambda/t$

$\hat{\lambda}/t$ is our estimate.