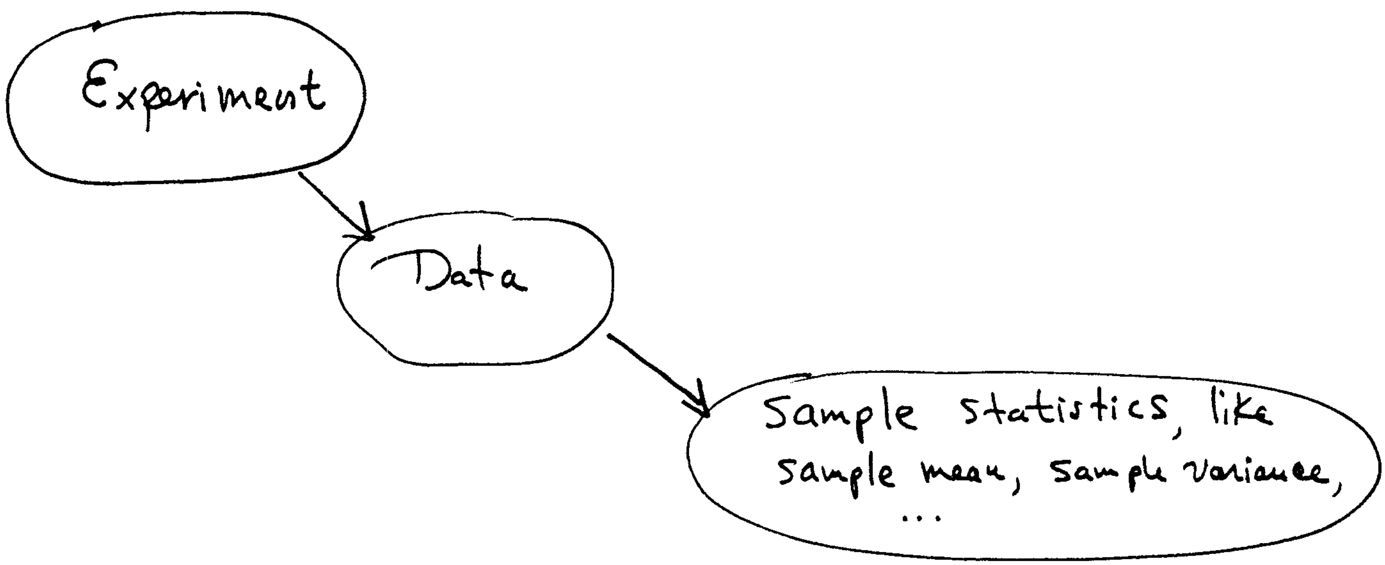
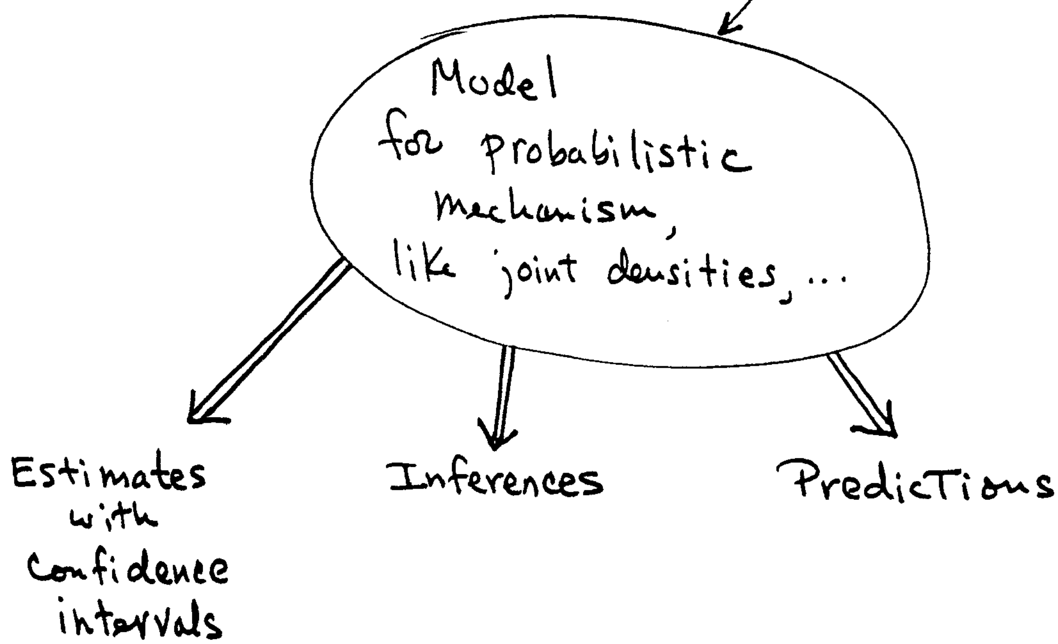
Interpreting Results [Lav82]

Descriptive Statistics



Inferential Statistics



Suppose x is a random variable with known prob. density fctn. $p(x)$.

$\boxed{\text{MEAN}}$ $$\mu = E(x) = \int_{-\infty}^{\infty} x \, p(x)\, dx \qquad \left(\begin{array}{c}\text{discrete} \\ \sum x_i p_i\end{array}\right)$$

$\boxed{\text{Variance}}$ $$\begin{aligned}\sigma^2 &= E\left[(x-\mu)^2\right] \\ &= E[x^2] - 2E[x\mu] + E[\mu^2] \\ &= E[x^2] - \mu^2\end{aligned}$$

σ is called <u>standard deviation</u>

Distinguish from <u>sample mean</u>, <u>sample variance</u>:

Suppose we have N independent observations of x, $x_1, x_2, \dots x_N$

- <u>Sample mean</u> $= \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$ $\left(\begin{array}{l}\text{a "Statistic"} \\ \text{a random Variable}\end{array}\right)$

$$E[\bar{x}] = \mu \qquad \underline{\underline{\text{unbiased}}}$$

- <u>Sample variance</u> $= s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \underbrace{\bar{x}}_{\text{sample mean}})^2$

why divide by $N-1$? $\rightarrow E[s^2] = \sigma^2$ $\underline{\underline{\text{unbiased}}}$

<u>Intuitively</u>, the N differences $(x_i - \bar{x})$ are not independent, because $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$

$\therefore$ there are only $(N-1)$ degrees of freedom in statistic

A proof is straightforward but requires some algebra

Some details, then:

<u>Lemma</u> Suppose samples $x_1, \ldots, x_N$ come from distribution with mean μ & variance σ^2, and are independent.

then $\boxed{E(\bar{x}-\mu)^2 = \frac{\sigma^2}{N}}$ $\rightarrow$ shows that standard deviation decreases as $1/\sqrt{N}$

<u>Proof</u> algebra

---

then
$$E[s^2] = E\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]$$
$$= \frac{1}{N-1} E\left[\underbrace{\sum x_i^2}_{R1.32} - N\bar{x}^2\right]$$
$$= \frac{1}{N-1}\left[N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)\right]$$
$$= \sigma^2 \quad \checkmark$$

A computational point:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Can compute by first finding $\bar{x}$, then direct evaluation.

But

$$s^2 = \frac{1}{N-1} \left[\sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2 \right]$$

$$= \frac{1}{N-1} \left[\Sigma x_i^2 - 2 (\bar{x})^2 N + N (\bar{x})^2 \right]$$

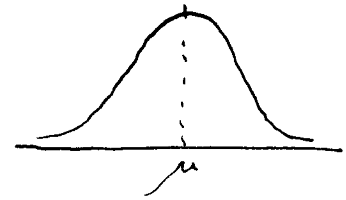
$$= \frac{1}{N-1} \left[\Sigma x_i^2 - N (\bar{x})^2 \right]$$

$$\boxed{s^2 = \frac{\Sigma x_i^2 - N(\bar{x})^2}{N-1}}$$

- fewer operations
- more accurate

Importance of Gaussian (Normal) Distribution

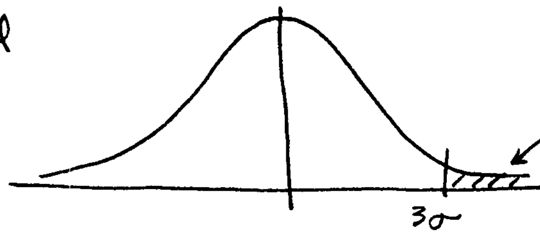
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



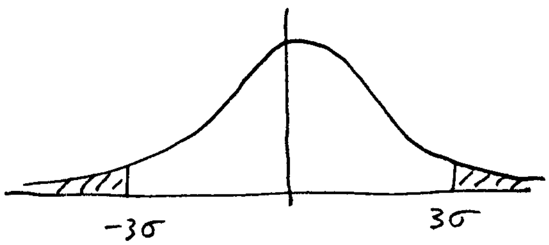
Often denoted $N(\mu, \sigma^2)$. $E(x) = \mu$ $\quad var(x) = \sigma^2$

We often deal with z = normalized Gaussian $N(0, 1)$

Tails are small

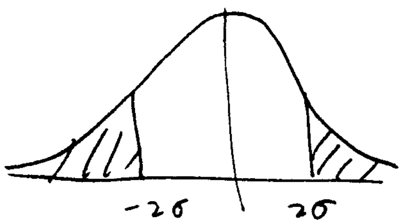


prob. = 0.00135



prob. of deviating more than 3σ from mean = 0.0027

prob. of not = 99.73%



& prob. of deviating less than 2σ = 95.45%

Why so Important?

Sum of independent observations converge to Gaussian under very general circumstances.

In nature, events that result from many small, independent effects tend to be Gaussion.

Central Limit Theorem

Suppose we sample $x_1, \ldots, x_n$ from a distribution with mean μ and variance σ^2.

Let
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{as usual}$$

then
$$\underbrace{z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}_{\text{Normalized random variable}} \longrightarrow N(0,1) \quad \text{Standard Normal}$$

Notice this is true for <u>any</u> parent distribution. (See probability books for technical details & conditions.)

Important Properties of Normal Distribution

1. Linear combination of Normals is also Normal
2. Normal has <u>maximum</u> <u>entropy</u> for given σ.
3. Least-squares becomes maximum likelihood
4. Many derived random variables have analytically known densities
5. Sample mean and variance of n identical, independent samples are <u>independent</u>; the sample mean is Normal $\underline{X}_n \sim N(\mu, \sigma/\sqrt{n})$.

Summary of Distributions of random Variables derived from Normal

Let x_i be n indep. ident. distributed samples from $N(\mu, \sigma)$.

SAMPLE MEAN • $\bar{x}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} x_i$ is distrib. as $N(\mu, \sigma/\sqrt{n})$

SAMPLE VARIANCE • $s_n^2 \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$

then

$$U = \frac{(n-1)\, s_n^2}{\sigma^2} \qquad \text{(Normalized)}$$

has a χ^2-distribution with $(n-1)$ degrees of freedom:

$$p(\chi^2) = \left[2^{n/2}\, \Gamma\left(\tfrac{n}{2}\right)\right]^{-1} (\chi^2)^{\frac{n}{2}-1}\, e^{-\chi^2/2} \qquad \chi^2 \geq 0$$

$E[U] = n-1$

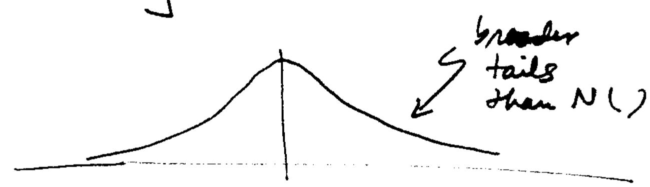
$\mathrm{var}[U] = 2(n-1)$

SAMPLE MEAN UNKNOWN VARIANCE • $\dfrac{\bar{x} - \mu}{s_n/\sqrt{n}}$ has a t-distribution

with $(n-1)$ degrees of freedom:

"Student-t" = W.S. Gosset

$$p(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{\pi n}} \frac{1}{\left[1 + t^2/n\right]^{\frac{n+1}{2}}}$$



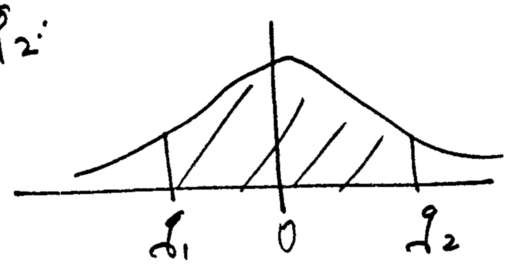
Confidence Intervals

Assume again $x_1, \dots, x_n$ i.i.d. Normal.

We want to know how far $\bar{x}_n$ might be from μ.

we know $\dfrac{\bar{x}_n - \mu}{S_n/\sqrt{n}}$ is Student-t distributed.

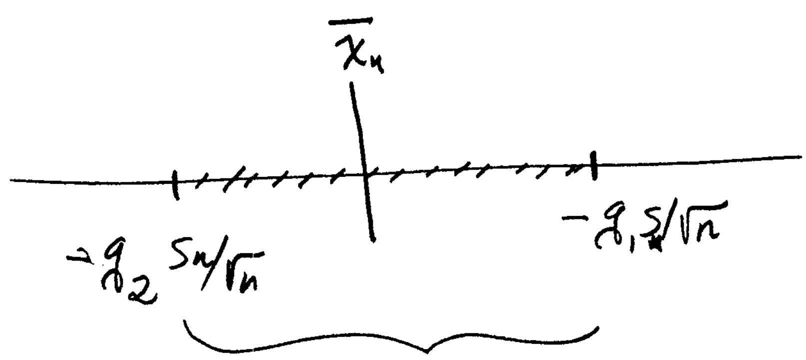
1. Pick $q_1, q_2 \ni$ student-t with $(n-1)$ degrees of freedom has 99% prob. (say) of lying between q_1, q_2:



$\boxed{\text{from TABLES}}$

2. $$\text{Prob.} \left\{ q_1 < \frac{\bar{x}_n - \mu}{S_n/\sqrt{n}} < q_2 \right\} = 0.99$$

$$\Rightarrow \text{prob.} \left\{ \bar{x}_n - q_2 \frac{S_n}{\sqrt{n}} < \mu < \bar{x}_n - q_1 \frac{S_n}{\sqrt{n}} \right\} = 0.99$$



"99% confidence interval"

~ prob. μ is here is 99%

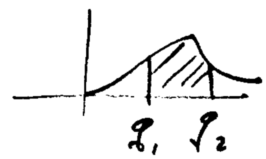
Confidence Interval for σ^2

Again, assume $x_1, \dots, x_n$ i.i.d. Normal.

We know $(n-1)\, s_n^2/\sigma^2$ is χ^2-distr. with $n-1$ D.O.F.

1. Pick $q_1, q_2 \ni$

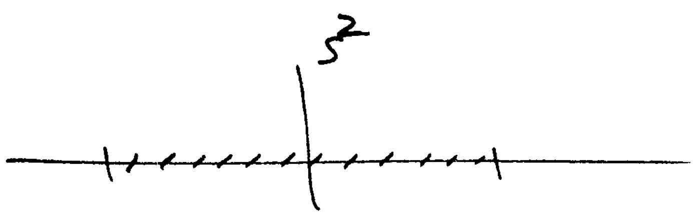
$$\text{prob.}\left\{ \chi^2_{\,n-1 \text{ D.O.F.}} \text{ between } q_1, q_2 \right\} = 0.99$$



2. $$\text{prob.}\left\{ q_1 < \frac{(n-1)s_n^2}{\sigma^2} < q_2 \right\} = 0.99$$

TABLES

$$\Rightarrow \text{prob}\left\{ \frac{(n-1)s_n^2}{q_2} < \sigma^2 < \frac{(n-1)s_n^2}{q_1} \right\} = 0.99$$

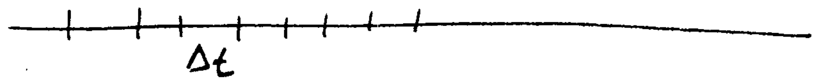


$$\underbrace{\frac{(n-1)s_n^2}{q_2} \qquad\qquad \frac{(n-1)s^2}{q_1}}_{\text{prob. that true } \sigma \text{ is here is } 99\%}$$

CAVEAT: Must be close to Normal to be valid

→ can often invoke Central Limit theorem to justify.

ADDITIONAL TOPIC: SIMULATING RANDOM ARRIVALS 1.3.10



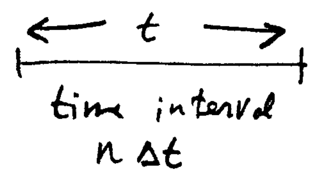
Assume prob. of arrival in any interval

- proportional to Δt, say $\mu \Delta t$, where μ = avg. # of arrivals/sec
- indep. of all other arrivals in other intervals
- small enough so we can neglect prob. of >1 arrivals.

then

prob. {no arrivals in intervals 1 to n}

$$= (1-\mu \Delta t)^n$$



$$= (1-\mu \Delta t)^{t/\Delta t}$$

Let $\Delta t \to 0$. $\lim (1-x)^{1/x} \to e^{-1}$, so

prob. {no arrivals till t} $= e^{-\mu t}$

prob. {first arrival before t} $= 1-e^{-\mu t}$

prob. density {time t to first event}

$$= p(t) = \frac{d}{dt}\left[1-e^{-\mu t}\right] = \mu e^{-\mu t}$$

"exponentially distributed" "Poisson arrivals" two Methods for generating such random arrivals:

I. take small Δt, flip coin with probability $\mu \Delta t$ each interval

II. Generate x uniform on $[0, 1]$, use $\boxed{-\frac{1}{\mu} \ln(x)}$ to determine next arrival

(See transformation method, p. 1.2.11)