Fundamentals of Statistical Analysis in Physics
Lectures 3, 4: Frequentist Statistical Analysis

African School of Physics 2022

Harrison B. Prosper

Kirby W. Kemper Endowed Professor of Physics
Department of Physics, Florida State University

29 November, 2022

Outline

Recap

- Collect raw data from a data generation mechanism.
- Process the raw data into meaningful observed data, D.
- Construct a statistical or probability model of all potentially observable data, X.
- Enter the observed data into the probability model thereby converting the latter into a likelihood function.
- Use the likelihood function to make inferences about the model.

Outline

Suppose there are only two possible outcomes of an experiment: a success or a failure.

Example (2.1 The LHC)

In n proton-proton collisions there are k successes, say the creation of a Higgs boson, and $n - k$ failures. A collision is an example of a trial.

What is the probability to get k successes out of n trials?

There is no answer!

Unless we make a sufficient number of assumptions.

Key Assumption

- The probability p of a success is the same for every trial.
- The order of trials is irrelevant.

Since there are only two possible outcomes, the probability of a success and a failure must add to one.

Therefore, the probability of a failure is $1 - p$.

Consequently, the probability of exactly k successes, which requires all other trials to fail, is

$$p_{k,n} = p^k (1-p)^{n-k}.$$

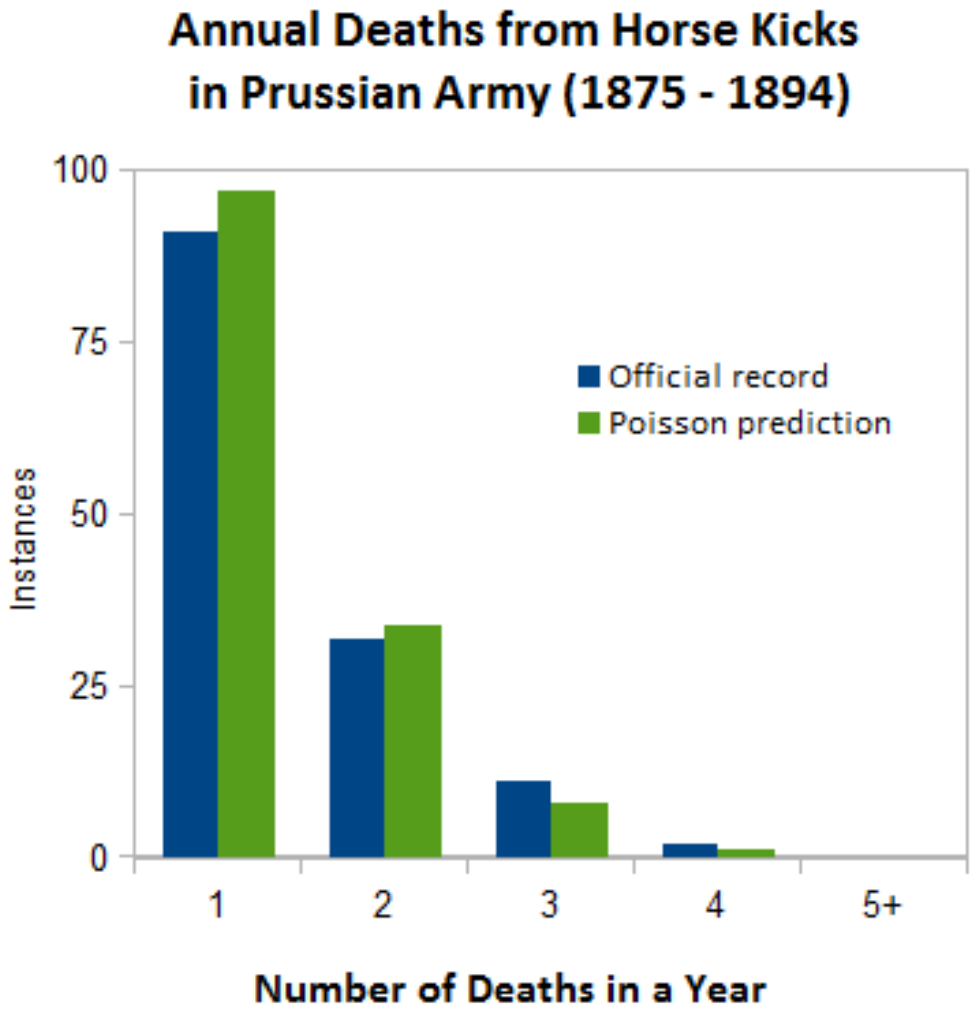
But there are $\binom{n}{k}$ ways to get exactly k successes in n trials.
We, therefore, conclude that the probability to get exactly k successes in n trials is given by the binomial distribution

$$\boxed{\text{binomial}(k; p, n) = \binom{n}{k} p^k (1-p)^{n-k}.}$$

The Poisson Distribution

In 1898, the Russian economist von Bortkiewicz published a book in which he presented data on the number of deaths per annum from horse kicks in the Prussian Army.

von Bortkiewicz noted that the distribution of observed counts could be modeled by the distribution first described (in 1837) by Siméon Poisson (1781 - 1840).



[1] https://mindyourdecisions.com/blog/2013/06/21/what-do-deaths-from-horse-kicks-have-to-do-with-statistics/

The Poisson Distribution

There are many situations in which the count n in the binomial distribution

$$\text{binomial}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k},$$

is large and the probability p of a success is small. Consider the limit $n \to \infty$ and $p \to 0$ while $\mu = pn$ and k remain fixed.

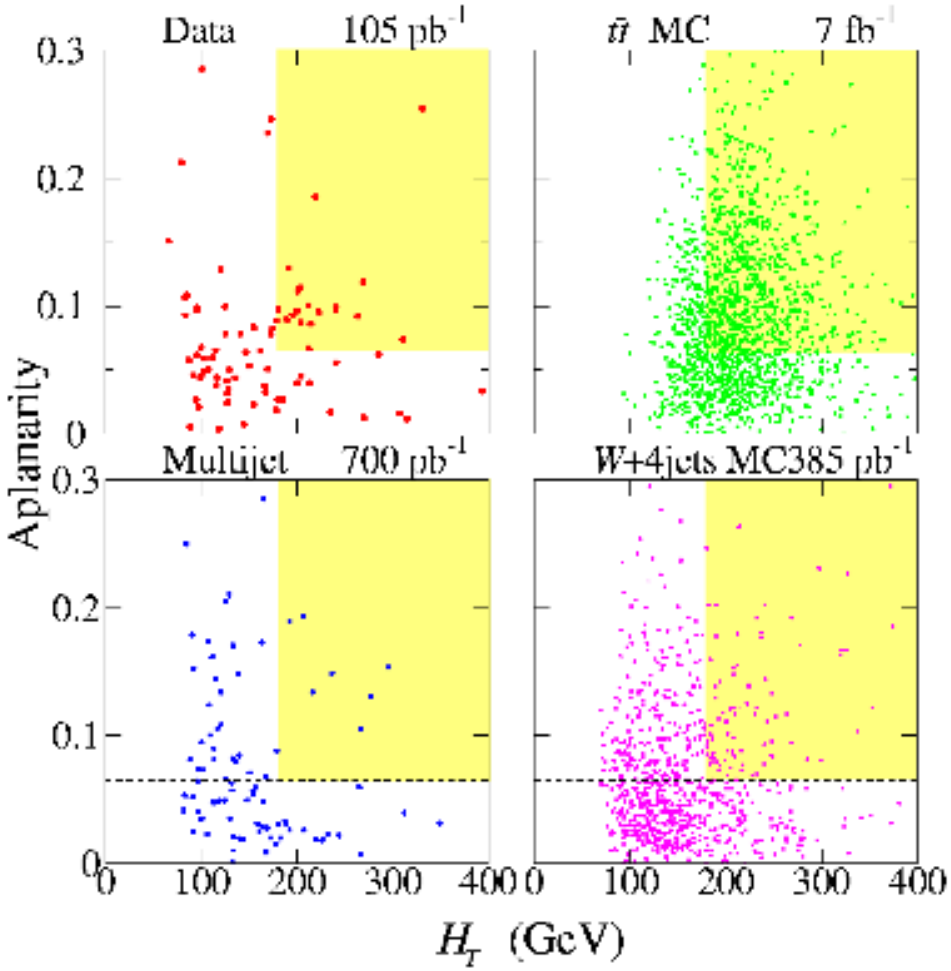
$$\begin{aligned}
\log[\text{binomial(k; n, p)}] &= \log n! - \log(n-k)! - \log k! \\
&\quad + \log p^k + \log(1-p)^{n-k}, \\
&\approx n \log n - n - (n-k)\log(n-k) + n - k - \log k! \\
&\quad + \log \mu^k - k \log n + \log(1-\mu/n)^{n-k}, \\
&= -\log k! + \log \mu^k - \mu,
\end{aligned}$$

yields the Poisson distribution,

$$\boxed{\text{Poisson}(k; \mu) = \mu^k \exp(-\mu)/k!}$$

Outline

DØ top quark discovery data (1995)



The DØ top quark discovery data.

We illustrate a few concepts in the frequentist approach by analyzing the DØ top quark discovery data:

$$N = 17 \quad \text{events},$$
$$B \pm \delta B = 3.8 \pm 0.6 \quad \text{background events}.$$

Step 1: Probability model

We take the probability (or statistical) model for the event count X to be

$$p(X|s, b) = (s + b)^X \exp(-(s + b))/X!,$$

where the parameters s and b are the (unknown) mean (top quark) signal count and mean background counts, respectively.

Step 2: Likelihood

By definition, the likelihood is just probability model into which the observed data, here we have a single datum N, has been inserted,

$$L(s, b) \equiv p(N|s, b) = (s + b)^N \exp(-(s + b))/N!.$$

To write down a likelihood for the background needs a bit more work.

Let's assume that the background estimate $B \pm \delta B = 3.8 \pm 0.6$ is the result of scaling down a count M by a factor k, that is, $B = M/k$.

Since, by assumption, M is a count, we may take its likelihood to be

$$p(M|kb) = (kb)^M \exp(-kb)/M!.$$

Moreover, since this is a Poisson distribution its standard deviation is $\sigma = \sqrt{kb} \approx \sqrt{M}$. This reasoning leads to the following ansatz

$$B = M/k,$$
$$\delta B = \sqrt{M}/k,$$

which can be inverted to arrive at

$$M = 40.11 \text{ events}$$
$$k = 10.56.$$

The non-integral value for $M = 40.11$ can be taken into account by writing $M! = \Gamma(M+1)$.

The overall likelihood for the data $D = N, M, k$ is, therefore,

$$\boxed{p(D|s,b) = \frac{(s+b)^N \exp(-(s+b))}{N!} \frac{(kb)^M \exp(-kb)}{\Gamma(M+1)}}.$$

Now that we have constructed our probability model, the next step is to measure the parameters, or to use the jargon of statisticians, estimate the parameters of the model.

To do this we must construct an estimator, that is, a procedure (typically a function) which, when data are entered into it, furnishes an estimate of the parameter(s) of interest.

Maximum Likelihood

In the method of maximum likelihood, used by Gauss and Laplace and systematically developed by Sir Ronald Fisher in the 1930s, estimators, $\hat{\theta}_i(X)$, are the solutions of the equations

$$S_i(\theta) \equiv \frac{\partial \log p(X|\theta)}{\partial \theta_i} = 0.$$

The functions $S_i(\theta)$ are called scores.

Let's apply this to our likelihood $p(D|s, b)$ for the parameter b, for a fixed, value of the mean signal count s. We find

$$\hat{b}(s) = \frac{g + \sqrt{g^2 + 4(1+k)Ms}}{2(1+k)}, \text{ where } g = N + M - (1+k)s. \quad (1)$$

When we insert the maximum likelihood estimate (MLE) of s, namely, $\hat{s} = N - M/k$ into $\hat{b}(s)$, we obtain the perfectly reasonable result $\hat{b}(\hat{s}) = M/k$.

Maximum Likelihood

The Good, the possibly Bad, and the perhaps Ugly of the maximum likelihood method.

The Good

- Maximum likelihood estimators (MLEs) are consistent.
- If an *unbiased* estimator for a parameter exists, the maximum likelihood procedure will find it.
- Given the MLE for θ, the MLE for $\mu = g(\theta)$ is just $\hat{\mu} = g(\hat{\theta})$.

The Bad?

- In general, MLEs are *biased*. Note, however, that a biased estimator may be more accurate than an unbiased one!

The Ugly

- Unfortunately, almost all the beautiful theorems about likelihoods are relevant asymptotically, that is, when the data are plentiful.

Step 3: Quantifying accuracy

Neyman argued that uncertainty about the value of a parameter μ should be quantified by constructing a confidence interval $[\,\mu_L(D), \mu_U(D)\,]$ for the parameter at a specified confidence level (CL).

For 1-parameter probability models, Neyman provided an algorithm that guarantees the construction of statements that satisfy the frequentist principle regardless of the true value of μ.

In our probability model, $p(D|s, b)$, the parameter of interest is $\mu = s$. Unfortunately, $p(D|s, b)$ also contains the nuisance parameter[2] b.

To proceed, we need to find a function, $t(s, b, D)$, of the parameters and the data such that the probability distribution $p(t)$ is independent of b.

A statistic $t(s, b, D)$ with this property is called a pivot.

[2] Recall that a nuisance parameter is simply a parameter that is not of current interest.

Step 3: Quantifying accuracy

In practice, we resort to the procedure below to find an *approximate* pivot $t(s, b, D)$. The statistic $t(s, b, D)$ is an approximate pivot in the sense that $p(t)$ still depends on b, albeit, we hope, weakly.

1. Find the (conditional) MLE for b, $\hat{b}(s)$, as a function of s (as we did 2 slides earlier).
2. Replace the nuisance parameter b in $p(D|s, b)$ by its conditional MLE $\hat{b}(s)$, thereby yielding a function called the profile likelihood $L_p(s) \equiv p(D|s, \hat{b}(s))$.
3. Compute $t(s) = -2 \log \lambda(s)$, where

$$\lambda(s) = \frac{L_p(s)}{L_p(\hat{s})},$$

is the profile likelihood ratio.

The procedure on the previous slide is motivated by the following theorem proved by Wilks[3] in 1938:

Wilks' Theorem

Consider the likelihood $p(x|\theta)$ with parameters $\theta = \theta_1, \ldots \theta_n$.

- Find the conditional MLEs for $n - k$ parameters and determine the profile likelihood, $L_p(\theta') = p(x|\theta')$, with k parameters θ'.
- Define the likelihood ratio $\lambda = L_p(\theta') \,/\, L_p(\hat{\theta}')$.

Then, in the large sample limit, $t = -2 \log \lambda$ is distributed as a χ^2_k variate with k degrees of freedom. (See χ^2 distribution in Appendix.)

This theorem holds provided certain conditions are met including 1) the estimates must not lie on the boundary of the parameter space and 2) every parameter is identifiable, that is, can always be estimated.

[3] S.S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, The Annals of Mathematical Statistics Vol. 9, No. 1 (Mar., 1938), pp. 60-62 (3 pages)

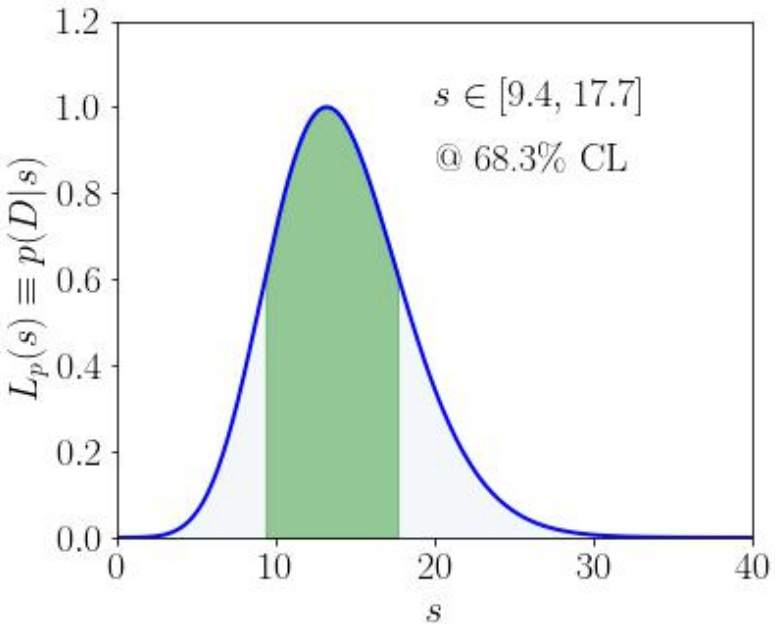
Step 3: Quantifying accuracy

According to Wilks' theorem,

$$t(s) = -2 \log \lambda(s) \approx \chi^2_1.$$

Therefore,
we can find an approximate confidence interval at 68% CL by solving

$$t(s) = 1,$$



which yields

$$[s_L(D), s_U(D)] = [9.4, 17.7].$$

In summary, our analysis of the DØ top discovery data, for which the signal estimate is $\hat{s} = N - M/k = 13.2$ events, yields the result

$$\hat{s} = 13.2^{+4.5}_{-3.8} \text{ events}.$$

Step 4: Quantifying the statistical significance of the signal

The profile likelihood ratio $\lambda(s) = L_p(s)/L_p(\hat{s})$, can be interpreted as the ratio of the "likelihood" of a given hypothesis H_s about the mean count s to the hypothesis $H_{\hat{s}}$ about the mean count $\hat{s}$ that best fits the data.

Furthermore, small values of $\lambda(s)$ and hence large values of $t(s) = -2\log\lambda(s) \approx \chi^2_1$ imply that hypothesis H_s is disfavored.

This gives us a way to test the null hypothesis H_0 that $s = 0$. If $t(0)$ is above an agreed-upon threshold, conventionally $t(0) \geq 25$ in particle physics, we agree to claim that the no-signal hypothesis is false and, therefore, we have made a discovery.

Given that $t(s) \approx \chi^2_1 = (\hat{s} - s)^2/\sigma^2$, we conclude that $Z \equiv \sqrt{t(0)} = \hat{s}/\sigma$ is the number of standard deviations by which the signal exceeds the background.

For the DØ top quark data our analysis yields $Z = 4.6\sigma$. The more sophisticated analysis by DØ takes this to the 5σ discovery threshold.

Outline

- Constructing a probability model is an important step in any non-trivial statistical analysis.
- Today, we sketched aspects of the frequentist approach to statistical inference by analyzing the DØ top quark discovery data.
- In the frequentist approach:
  - the method of maximum likelihood is a general method for estimating the parameters of a probability model.
  - the profile likelihood is the standard (approximate) way to deal with nuisance parameters.
  - provided that certain conditions are met, Wilks' theorem is the basis of a standard (approximate) method to quantify uncertainty.
- Extensions of Wilks' theorem exist for situations in which the relevant conditions are not met, but discussions of these extensions are beyond the scope of these lectures.

Outline

χ^2 Distribution

Write $z = (x - \mu)/\sigma$, where $x \sim \mathrm{Gauss}(\mu, \sigma)$[4] and consider the sum

$$t = \sum_{i=1}^{n} z_i^2.$$

What is the probability density function (pdf) of t? For any well-behaved probability density function, $p(z_1, \cdots, z_n)$, the pdf of t, $p(t)$, is given by the random variable theorem[5]

$$\boxed{p(t) = \int dz_1 \cdots \int dz_n \, \delta\left(t - g(z_1, \cdots, z_n)\right) \, p(z_1, \cdots, z_n)},$$

where $g(z_1, \cdots, z_n)$ is the function, such as the sum above, that maps z_1 to z_n to t. The δ-function imposes the constraint $t = g(z_1, \cdots, z_n)$.

[4] the symbol $\sim$ in this context means "sampled from".

[5] *A theorem for physicists in the theory of random variables*, D. Gillespie, Am. J. of Phys. **51**, 520 (1983).

χ^2 Distribution

First note that $p(z_1, \cdots, z_n) = p(z_1)p(z_2)\cdots p(z_n)$ and

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega x}\, d\omega.$$

Putting together the pieces and shuffling the order of integration, we get

$$p(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} \prod_{j=1}^{n} \int_{-\infty}^{\infty} e^{-i\omega z_j^2}\, p(z_j)\, dz_j.$$

$$
\begin{aligned}
p(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} \prod_{j=1}^{n} \int_{-\infty}^{\infty} e^{-i\omega z_j^2}\, p(z_j)\, dz_j, \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} \left(\int_{-\infty}^{\infty} e^{-i\omega z^2}\, \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \right)^n, \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-(2i\omega+1)z^2/2}}{\sqrt{2i\omega+1}}\, d\sqrt{2i\omega+1}z \right)^n, \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{e^{i\omega t}}{(2i)^{n/2}}\, \frac{1}{(\omega - i/2)^{n/2}}.
\end{aligned}
$$

χ^2 Distribution

$$p(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, \frac{i e^{i\omega t}}{(2i)^{n/2}} \, \frac{1}{(\omega - i/2)^{n/2}}.$$

We can compute the integral above using the residue theorem

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, F(\omega) \, \frac{1}{(\omega - \omega_0)^n} = \lim_{\omega \to \omega_0} \frac{1}{(n-1)!} \frac{d^{n-1}}{d\omega^{n-1}} F(\omega),$$

for pole singularities that do not lie on the real line.

Well, we could if it wasn't for a singularity that involves an annoying square-root!

No matter, we avail ourselves of a time-honored strategy of physicists: solve a simpler problem then generalize its solution by inspection!

χ^2 Distribution

Writing $m = n/2$, and performing the integral for integer m, we find

$$\frac{1}{2\pi i}\int_{-\infty}^{\infty} d\omega\, \frac{i e^{i\omega t}}{(2i)^m}\, \frac{1}{(\omega - i/2)^m} = \frac{1}{\Gamma(m)}\frac{t^{m-1}\, e^{-t/2}}{2^m}.$$

This result remains valid for non-integral values of m. Therefore, the pdf of the sum of n standardized Gaussian variates is ($t = \chi^2$)

$$\boxed{p(t) = \frac{1}{\Gamma(n/2)}\frac{t^{n/2-1}\, e^{-t/2}}{2^{n/2}},\ \text{mean } n,\ \text{variance } 2n}$$

.

Cauchy Distribution

Let $x, y \sim \text{Gauss}(0,1) \equiv g(x)$. What is the pdf of $t = y/x$?
It is given by

$$\begin{aligned} p(t) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, g(x)\, g(y), \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, e^{-\frac{1}{2}(x^2+y^2)}. \end{aligned}$$

This integral is begging us to use polar coordinates, $y = r \sin\theta$, $x = r\cos\theta$ and $dx\, dy \to r\, dr\, d\theta$, so that we can write

$$\begin{aligned} p(t) &= \frac{1}{\pi} \left(\int_0^{\infty} e^{-\frac{1}{2}r^2}\, r\, dr/2 \right) \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta, \\ &= \frac{1}{\pi} \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta. \end{aligned}$$

At first glance, the odd looking beast

$$p(t) = \frac{1}{\pi} \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta,$$

looks tricky! But, recall that $\delta(h(\theta)) = \delta(\theta - \theta_0)/|dh/d\theta|$, where θ_0 is the root of $h(\theta)$.

For this problem, $h(\theta) = t - \tan\theta = 0$ and $1/|dh/d\theta| = \cos^2\theta$. Therefore,

$$\begin{aligned} p(t) &= \frac{1}{\pi} \int_0^{2\pi} \delta(\theta - \theta_0)\, \cos^2\theta\, d\theta, \\ &= \frac{1}{\pi}\, \cos^2\theta_0. \end{aligned}$$

But, $\tan\theta = t \implies \cos\theta = 1/\sqrt{1+t^2}$. Therefore,

$$\boxed{p(t) = \frac{1}{\pi(1+t^2)}}$$